MAT 8480 - Data Mining Course Project

December 18th, 2019

Daniel Preston, Zackary Scayler, Aaron Siegel, Sebastian Tilson

aka ‘Lift Maximizers’

Table of Contents

Section 1 - Introduction

The following report covers an exploratory analysis of the dataset, clustering, and predictive modeling exercises to predict which customers are most likely to book another tour in the next 12 months. In order to understand the dataset and prepare for the modeling phase, we inspected each variable using graphical and numerical techniques. After performing this inspection, we consulted with each other as to how best to clean the data and prepare it for the modeling phase. This cleaning and preparation included marking missing observations for each variable, fixing categories that were mis-labeled, consolidating categories for categorical variables (using our assumptions or decision trees), and changing variable types. We also created a few new variables where necessary, which we will explain further. After all variables were cleaned, we combined them into a final dataset, calculated measures of variable importance, and split the data (50-25-25) to prepare for training and scoring. Following these essential preparatory steps, we dive into clustering and predictive modeling. We iterate through logistic regression, decision tree, random forest, and an artificial neural network model. By the end of the exercises, we select a random forest model which we interpret with a decision tree. We use the f-score as our performance metric and find that most models land around the 0.54 mark.

Section 2 - Data Exploration, Data Cleaning, Variable Transformation, Variable Imputation

Measures of Variable Importance

Below is a table of variable importance. It ranks all variables (besides Target and ID variables) by their ROC measures. The ROC Index represents the power of a given input variable to help distinguish between the target variable. In the Appendix, there are two supplemental tables, one for categorical variables and the other for numeric variables, which provide additional measures of variable importance like the chi-squared statistic and t-statistic.

Variable Name	ROC	ROC Rank
Past_Trips	0.647769	1
Email	0.632545	2
DB_Enter_Months	0.626542	3
Outbound_Intr_Gateway	0.580477	4
Overall_Impression	0.552666	5
Pre_Departure	0.548776	6
SourceType	0.544732	7
Reference	0.538506	8
Pax_Category	0.538262	9
TourCode	0.536383	10
Grp_Size	0.536157	11
Book_Months	0.532887	12
TravelAgain	0.531312	13
Excellent_Hotels	0.529584	14
Excellent_Optionals	0.528955	15
Hotels_Avg	0.527509	16
Extension	0.524135	17

Excellent_GUSS	0.523271	18
Meals_Avg	0.523127	19
Outbound_Domestic_Gateway	0.522956	20
Return_Domestic_Gateway	0.522624	21
Main_Ext	0.522311	22
Return_Intr_Gateway	0.522072	23
Grp_Size_Cat	0.5216	24
Recommend_GAT	0.521452	25
Groups_Interest	0.521144	26
Capacity	0.520685	27
Optionals	0.519979	28
Return_Connect_Time_Mins_2	0.519305	29
Domestic_Depart_Time_AM	0.519121	30
State	0.518927	31
Insurance	0.518641	32
Tour_Days	0.51826	33
Intr_Arrival_Time_AM	0.517937	34
Bus_Avg	0.517079	35

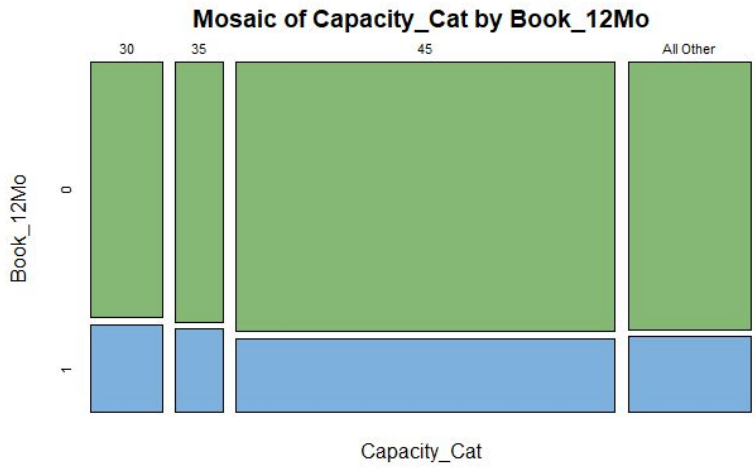
Flight_Itin	0.516945	36
Good_Hotels	0.51679	37
Intr_Depart_Time_AM	0.516236	38
Meals_2orUnder	0.515969	39
Hotel_2orUnder	0.515627	40
Tour_Type	0.515317	41
Excellent_Meals	0.515281	42
Return_Connections	0.514525	43
Capacity_Cat	0.513759	44
GUSS_Avg	0.513324	45
Poor_Meals	0.512516	46
Outbound_Connect_Time_Mins_1	0.511918	47
Good_Buses	0.51106	48
Fair_Meals	0.510992	49
Fair_Hotels	0.51051	50
Return_Connect_Time_Mins_1	0.510178	51
Excellent_Buses	0.508297	52
FY	0.508106	53

Poor_Hotels	0.508015	54
Start_Day	0.507607	55
Total_Outbound_Connect_Time	0.50716	56
TourPriceCat	0.507148	57
Good_Meals	0.506988	58
GUSS_2orUnder	0.506745	59
Age	0.506196	60
Outbound_Connections	0.505993	61
FltGty	0.505808	62
Total_Return_Connect_Time	0.505618	63
Voucher_Event	0.505167	64
Fair_GUSS	0.504798	65
Good_Optionals	0.504538	66
Fair_Optionals	0.504477	67
End_Day	0.504229	68
Optionals_2orUnder	0.504029	69
Tour_Season	0.503908	70
Good_GUSS	0.503777	71

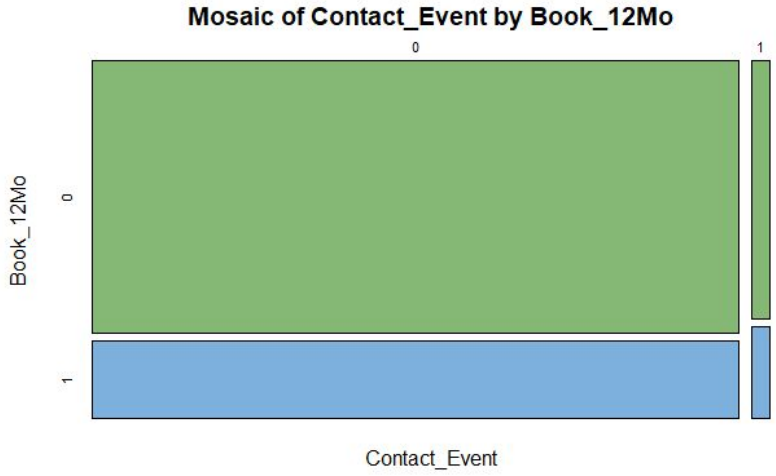
Contact_Event	0.503626	72
Eval_Contact_Days	0.503583	73
Outbound_Connect_Time_Mins_2	0.503409	74
Poor_GUSS	0.503068	75
TourDate_WeekYear	0.502882	76
Bus_2orUnder	0.501853	77
Complaint_Event	0.501443	78
Fair_Buses	0.501401	79
Optionals_Avg	0.501364	80
Tour_Region	0.501242	81
Poor_Optionals	0.501002	82
TD_Overall	0.500895	83
Poor_Buses	0.500399	84
Domestic_Arrival_Time_AM	0.500077	85
Promo_Disc	0.500055	86

Generated Variables

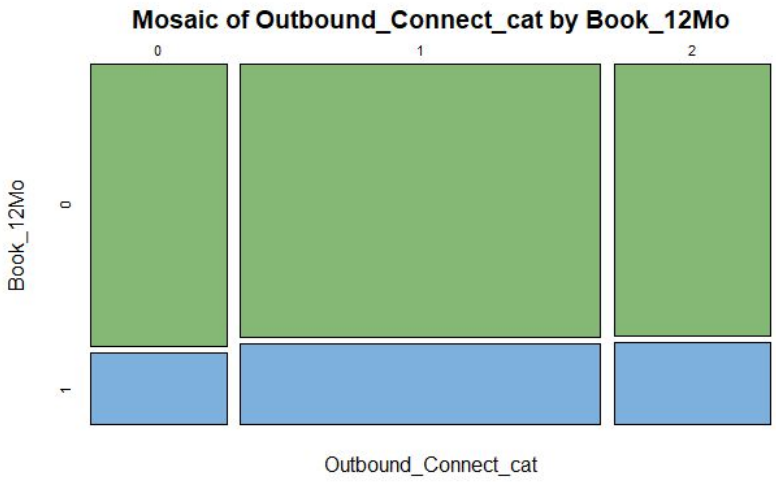
- **Capacity_Cat:** Tour capacity consolidated into three different categories, as shown below.



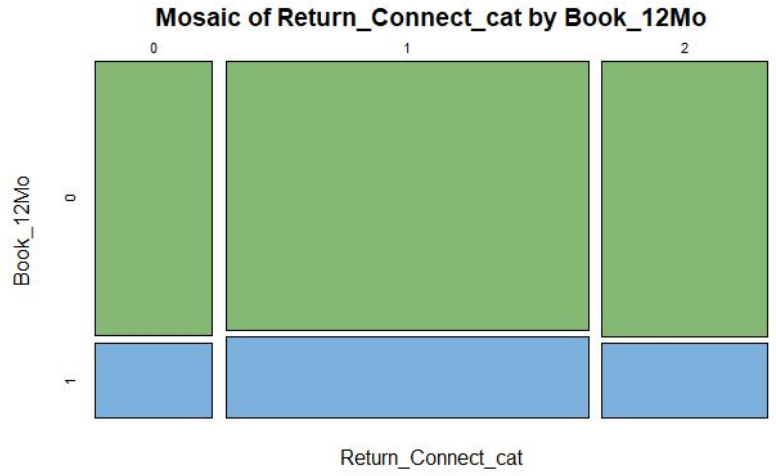
- **Contact_Event:** If the number of days after which the company made contact with the customer after they complained on the evaluation form is greater than 0, we assumed a “contact event” took place.



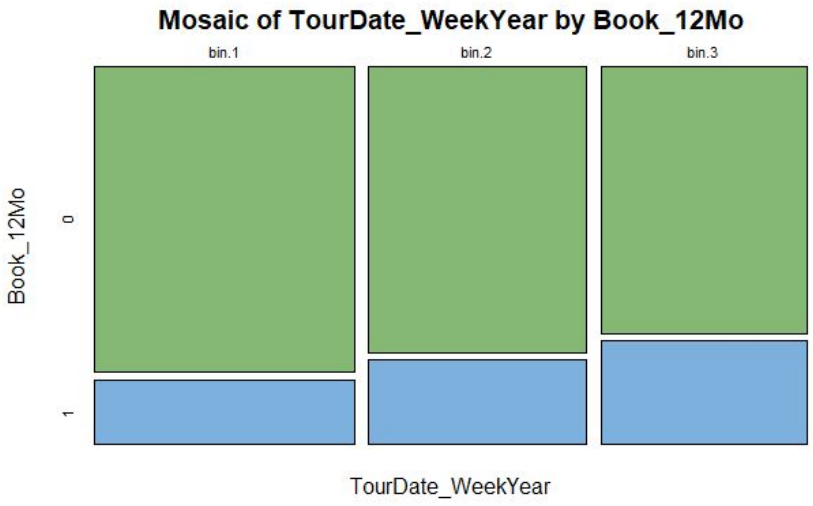
- **Total_Outbound_Connect_Time:** The sum of two outbound connect times.
- **Total_Return_Connect_Time:** The sum of two return connect times.
- **Outbound_Connect_cat:** Addressing errors in Outbound_Connections, by defining level 0 for no connection flight code,1 for one connection flight code, and 2 for two flight codes in data. Where level "2" represents two or more connections in Outbound_Connections.



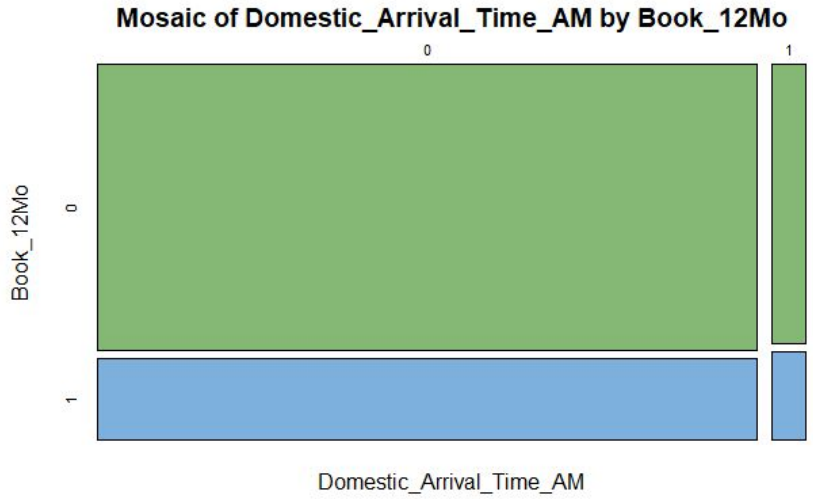
- **Return_Connect_cat:** Addressing errors in Return_Connections, by defining level 0 for no connection flight code, 1 for one connection flight code, and 2 for two flight codes in data. Where level "2" represents two or more connections in Return_Connections.



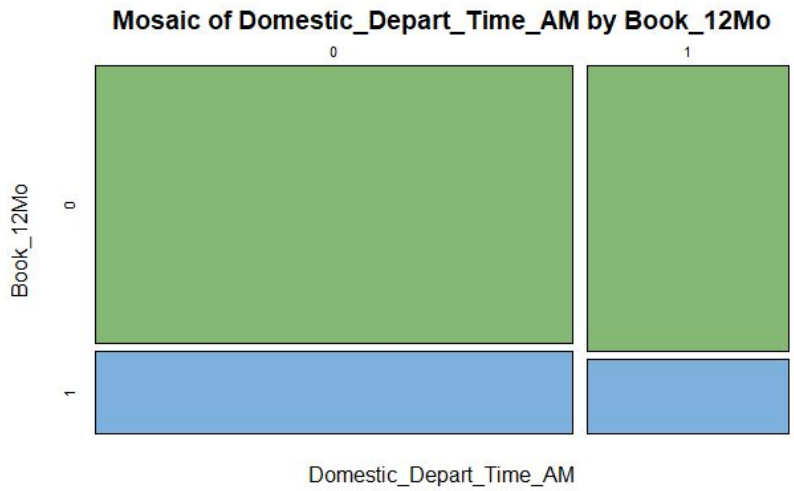
- **TourDate_WeekYear:** The week and year the tour took place, derived from TourDate variable and binned using tree.bins ANOVA method with cp parameter 0.003.



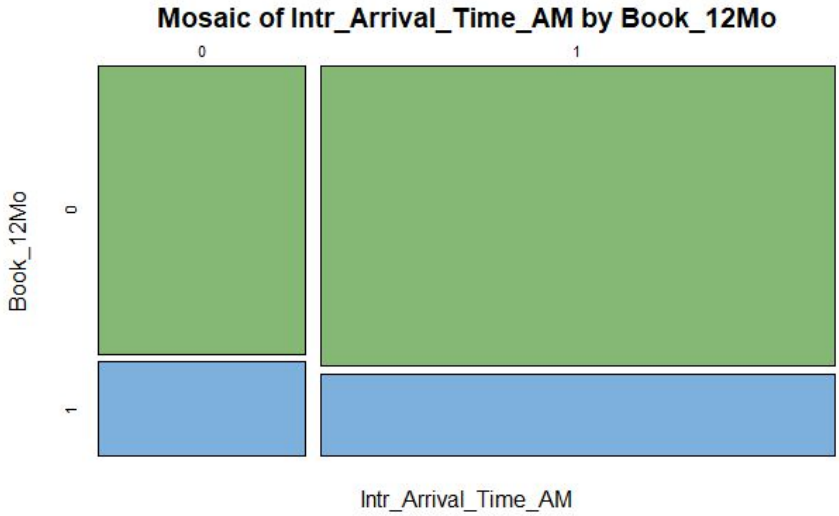
- **Domestic_Arrival_Time_AM:** Whether the local time at which the customer arrived was in the morning.



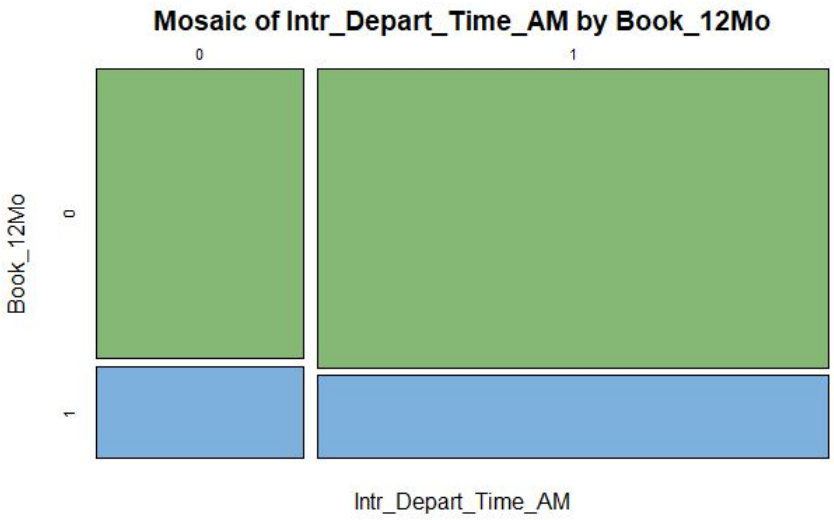
- **Domestic_Depart_Time_AM:** Whether the local time at which the customer departed was in the morning.



- **Intr_Arrival_Time_AM:** Whether the local arrival time of the tour start city that the customer is flying into is in the morning.



- **Intr_Depart_Time_AM:** Whether the local time that the customer had to depart at is in the morning.



Rejected Variables

- **TourDate:** We transformed this variable to TourDate_WeekYear to make it less unwieldy.
- **Domestic_Arrival_Time:** This variable was transformed into a binary (**Domestic_Arrival_Time_AM).**
- **Domestic_Depart_Time:** This variable was transformed into a binary (**Domestic_Depart_Time_AM).**
- **Intr_Arrival_Time:** This variable was transformed into a binary (**Intr_Arrival_Time_AM).**
- **Intr_Depart_Time:** This variable was transformed into a binary (**Intr_Depart_Time_AM).**
- **Promo_Disc:** Due to near zero variance.

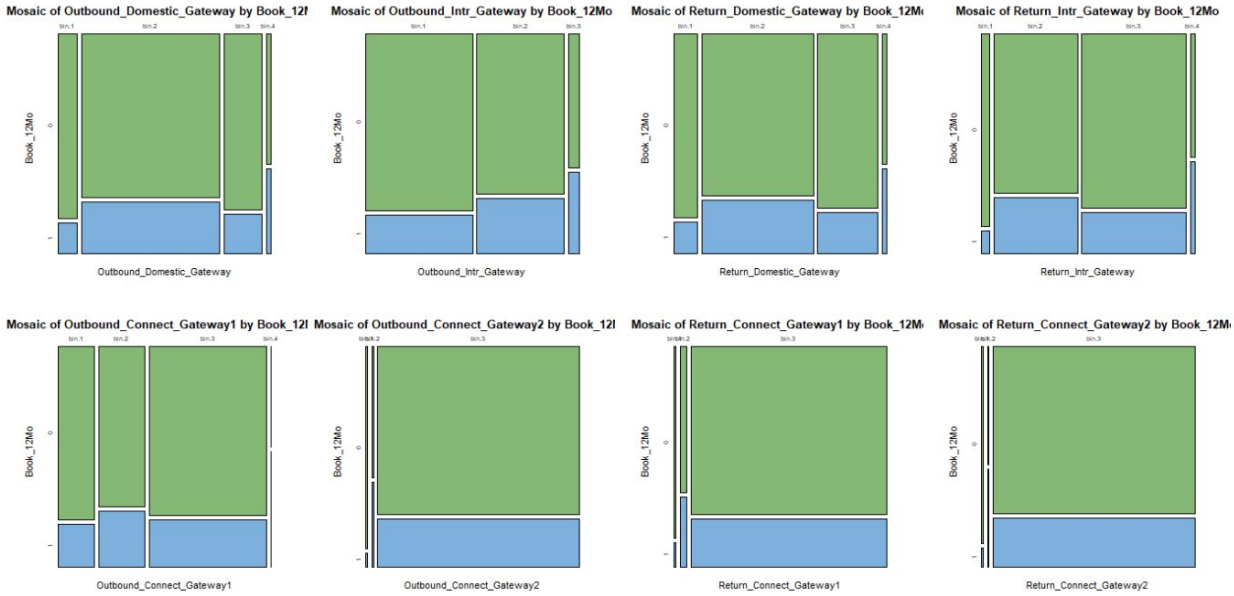
Identifier Variables

There are five identification variables which we will not use during the modeling phase of the project, but are important for understanding a potential unit of observation for the analysis. These five variables are ***EvalID***, which is a sequential number that is used to identify all the variables that are associated with evaluations. Not surprisingly, there is a customer identifier from the tour company's CRM system (***Cus_ID***). There are 21,702 unique customer identifiers. Additionally, there are 1,390 unique ***ProdTour_ID***, which is a tour number from the operational system, and 2,116 unique values of ***SalesTourID*** which is similar to ***ProdTour_ID*** but has unique values for the main tour and extension tour. Lastly, there is a ***Trip_No***, which is a tour key from the sales system that is unique to a customer's trip. The data dictionary explains that different customers going on the

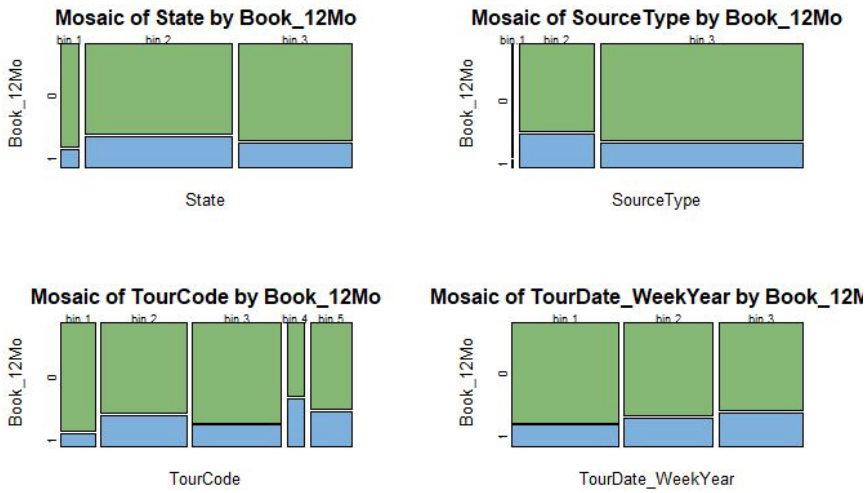
same trip will have the same ***SalesTour_ID*** but different ***Trip_No***. There are 23,459 unique values of ***Trip_No***. All five of these variables helped tie together data from different databases at the company into this convenient flat file we were offered.

Variables with Categories Consolidated Using Decision Trees

Eight category variables (Outbound_Domestic_Gateway, Outbound_Intr_Gateway, Return_Domestic_Gateway, Return_Intr_Gateway, Outbound_Connect_Gateway, Outbound_Connect_Gateway2, Return_Connect_Gateway1, and Return_Connect_Gateway2) have many different categories, the vast majority of which are not listed in the data dictionary. To consolidate the categories for these variables, which capture airport codes, we implemented decision trees to put each airport code for each variable into one of several bins. To show how this consolidation process was implemented, we included a table in the Appendix to show the binning pattern for one of the variables (Outbound_Domestic_Gateway). Below is the Mosaic plot for each variable.



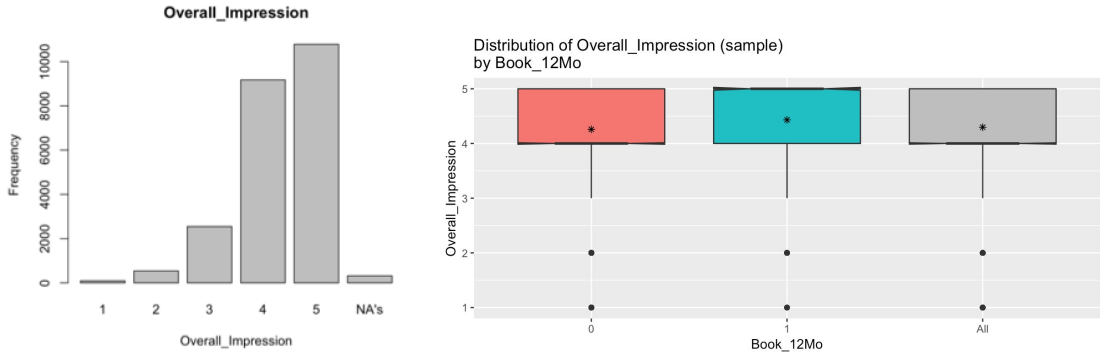
Three category variables (State, SourceType, and TourCode) had many categories not defined in the data dictionary, some of which contained arbitrary or conflicting labels. One generated variable (TourDate_WeekYear) also had too many categories to be useful as a predictor in the modeling process. Decision trees were implemented in the same manner as the above variables, using a 50% training data split, to consolidate these categories into a smaller number of bins. Below is the Mosaic plot for each variable.



Ratings of Trip Components

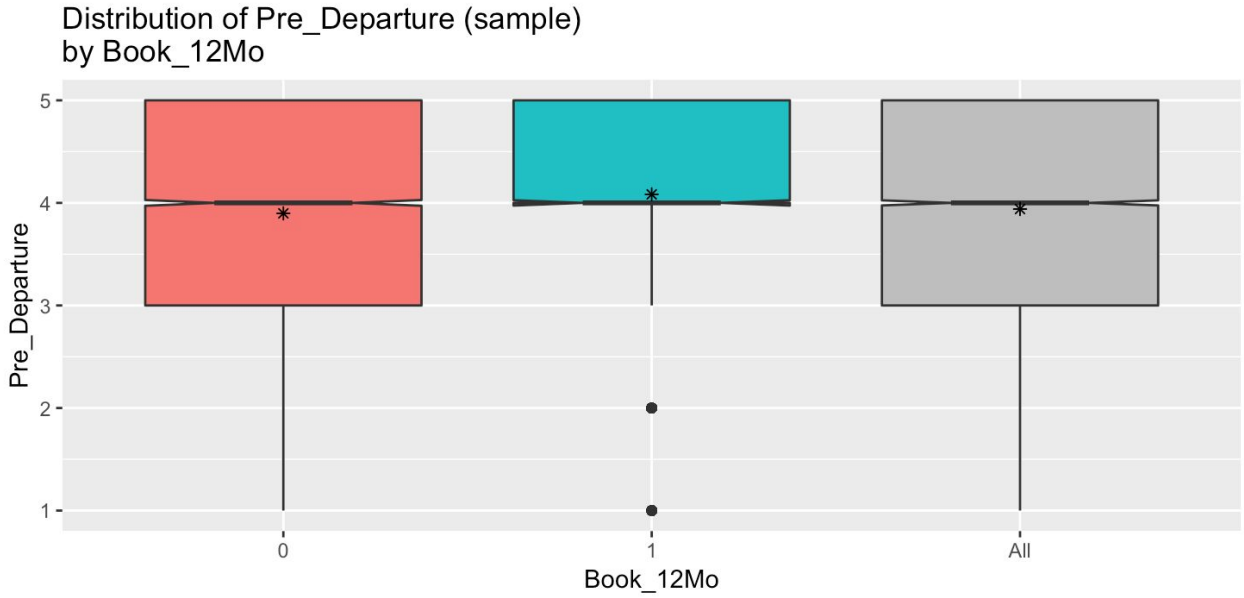
A significant number of variables measure ratings that customers gave about various aspects of their trip. Below, we provide explorations of the averages of these categorical variables like Overall_Impression, Pre_Departure, Flight_Itin, TD_Overall, Hotels_Avg, Meals_Avg, GUSS_Avg, Optionals_Avg, and Bus_Avg. There are also many variables that measure counts of particular ratings, for instance Meals_2orUnder. We do not provide an exploration of each and every one of these variables because they are almost all low on variable importance measures, and all share a similar extremely right skewed distribution.

- **Overall_Impression**: What was your overall impression of the tour? 1-Poor to 5-Excellent. We converted the 0 ratings to missing values. Overall_Impression may not seem to be strongly correlated with rebooking in 12 months but it rates highly on variable importance measures.

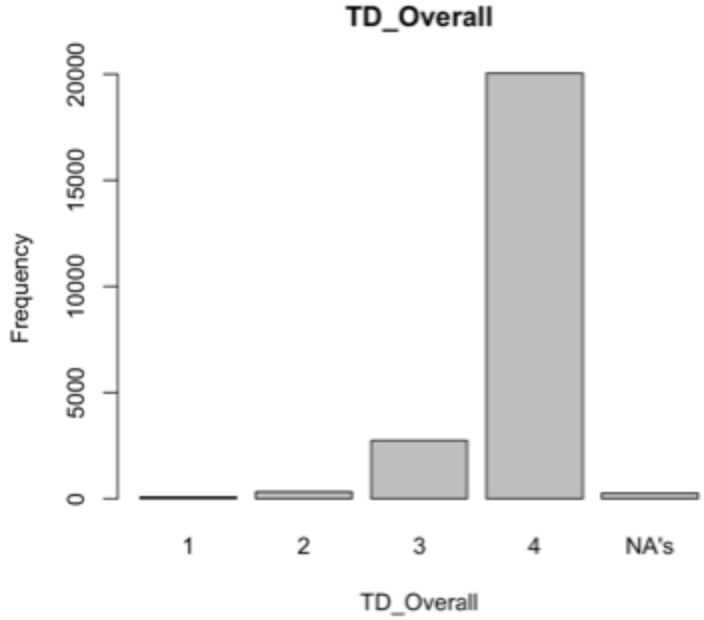


- **Pre_Departure**: Quality of Service Prior to Departure? 1-Poor to 5-Excellent. We converted the 0 ratings to missing values. Pre_Departure may not seem to be

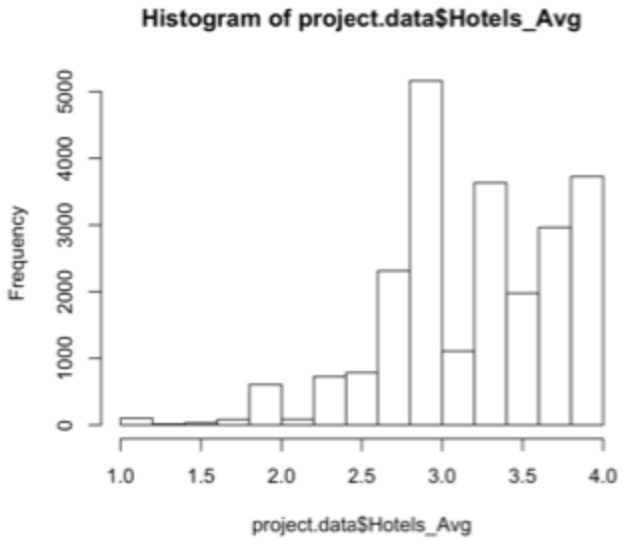
strongly correlated with rebooking in 12 months but it rates relatively highly on variable importance measures.



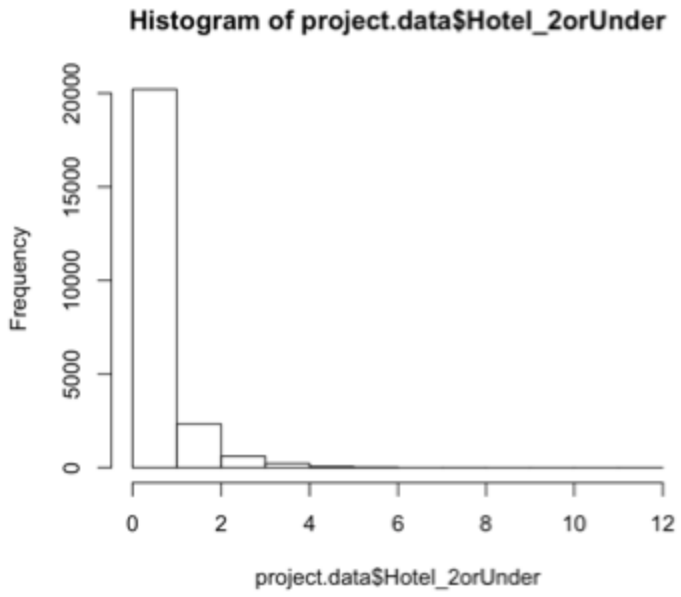
- **Flight_Itin:** Rating of flight itinerary/routing (1-poor to 5-excellent). Values of 0 were converted to missing values. This variable does not seem to be particularly associated with rebooking within 12 months.
- **TD_Overall:** Rating of tour director (1-poor to 4-excellent). Values of 0 were converted to missing values. This variable does not seem to be particularly associated with rebooking within 12 months, perhaps because the vast majority of tour directors received a rating of 4.



- **Hotels_Avg:** The average ratings of all hotels on the tour (1-poor to 4-excellent) and 0 ratings were converted to missing values. This variable seems to be somewhat related to the target variable of rebooking within 12 months.



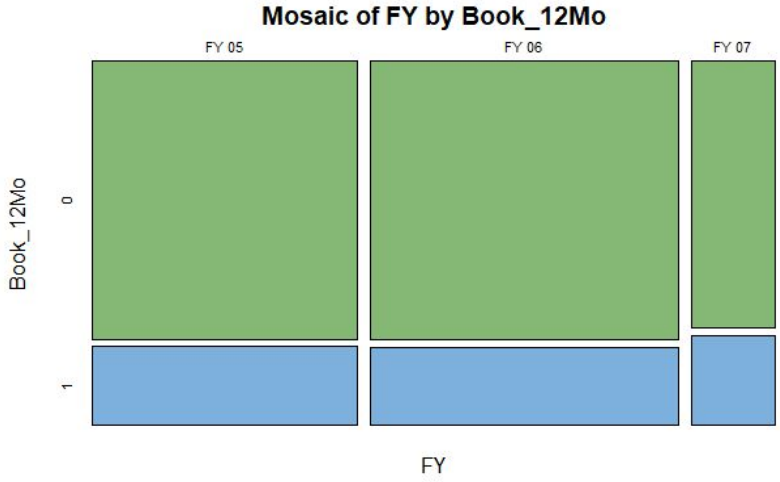
- The remainder of the trip component ratings have similar distributions to the variables above, with left skew. For these variables of averages, if the average was 0, we assumed it to be missing because that suggests all of the individual ratings were zero. One variable worth noting is the Bus_Avg variable. More than half of the observations had a missing value so instead of imputing it, we defined a new variable that is 0 or 1 depending on whether Bus_Avg is missing or not.
- As an example of one of the variables that captures counts of particular ratings, here is the variable **Hotels_2orUnder.** Like many of the other counts of certain ratings, it is very right skewed and ranks low on measures of variable importance.



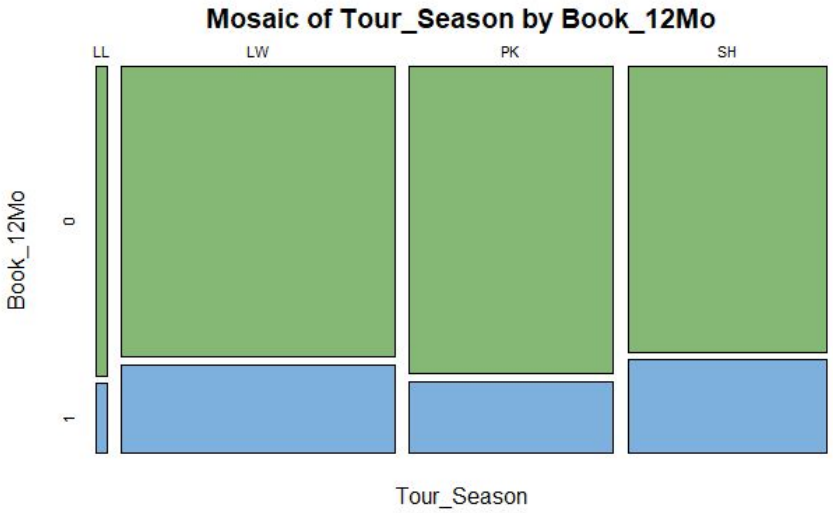
These variables, relating to customer survey data, were used to fit a Gaussian Mixture Model to create a classification cluster as described in Section 3.

Remaining Variables

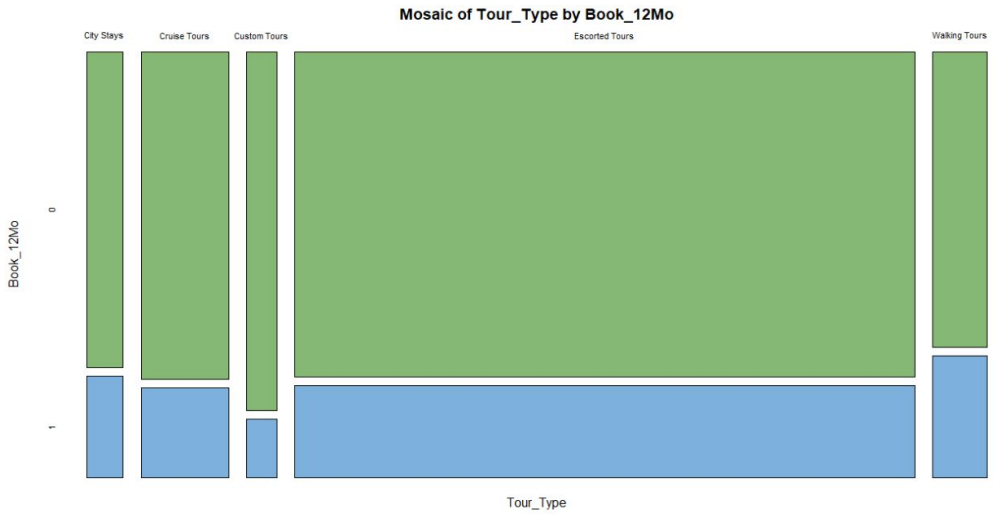
- **Fiscal Year (FY):** This variable indicates the fiscal year the trip was taken. We can see below that there are 23,459 values for FY and no records are missing. As expected, there are 3 distinct values and nearly half of the tours were taken in fiscal year 2006. We can see below the mosaic plot for FY against the target variable, Book_12Mo. It doesn't appear that the FY is strongly related to whether or not a customer will book another tour within the next 12 months.



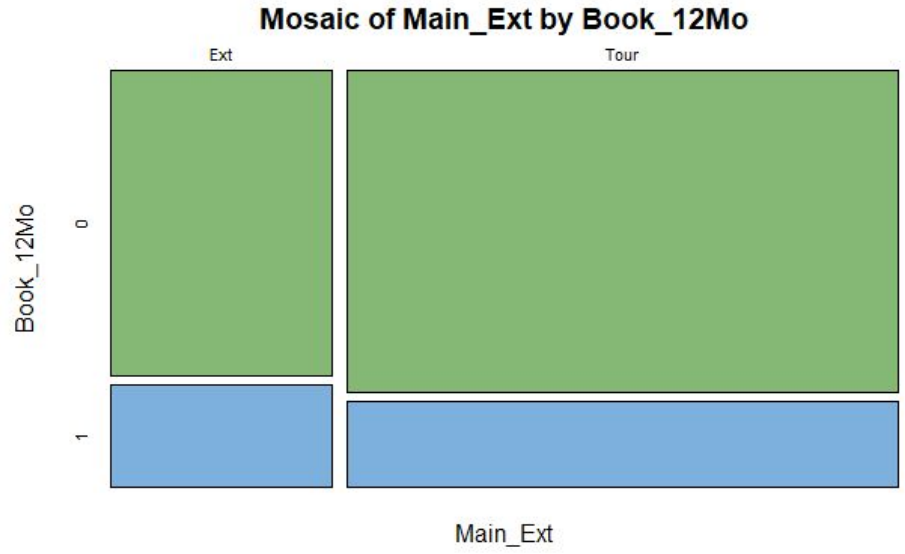
- **Tour_Season:** This variable tells you the season by geographic area that the tour is going to. For example, a tour to Italy in January would be LW (Low Season), where as a tour to Australia in January would be PK(Peak Season). There are 23,459 values, 0 records are missing, and there are 4 distinct values as expected. Below we can see that the relationship between booking another tour within the next 12 months and the tour season isn't strongly related, but Low Season travel tends to be the most crowded. This does not make good business sense. Moreover, there are two levels defined in the variable description, ML (Mid Low Season) and HS (High Shoulder Season), that are not observed, and one level LL that is not defined.



- **Tour_Type:** This variable indicates the type of tour the customer went on. There are 23,459 values, 0 are missing, and as expected there are 5 distinct values. Nearly 3 in 4 customers went on an escorted tour. It appears from the mosaic plot that those that went on walking tours and city tours were more likely to book again.



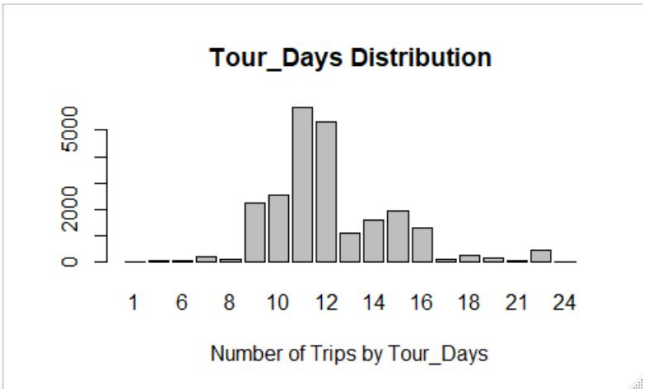
- **Main_Ext:** This variable indicates the type of tour. A value of ‘tour’ is a main tour and ‘ext’ is an extension to the main tour. There are 23,459 values, 0 missing values, and as expected 2 distinct values. Over 7 in 10 customers only took the main tour. Judging by the mosaic chart below, it appears that customers who took an extension tour were slightly more likely to book another tour within 12 months compared to those who only took a main tour.



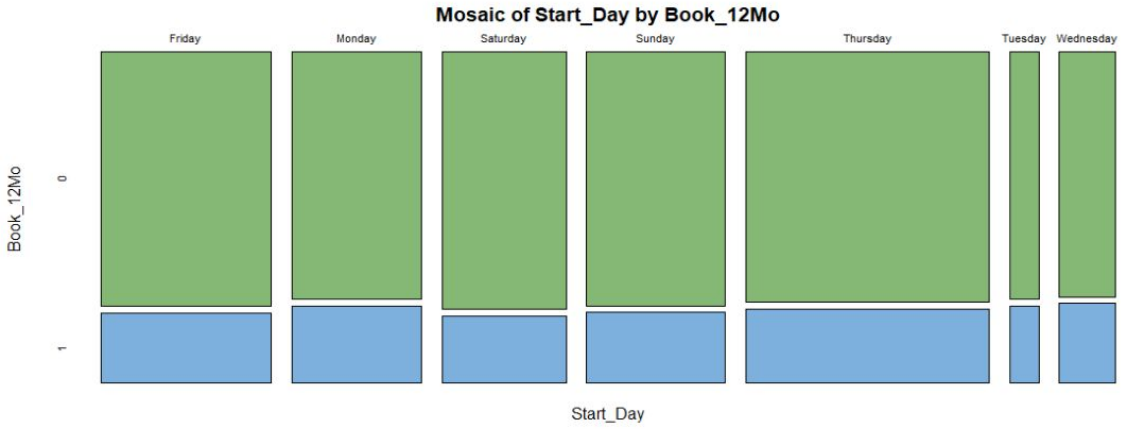
- **Tour_Days:** This variable indicates the number of days of the tour. There are 23,459 values, 0 missing values, and 19 distinct values. This is a numeric variable. The 25th percentile is 11, while the 75th percentile is 14. The Point-Biserial Correlation between Tour_Days and Book_12Mo is only -0.027,

indicating no relationship between this potential input variable and the target variable.

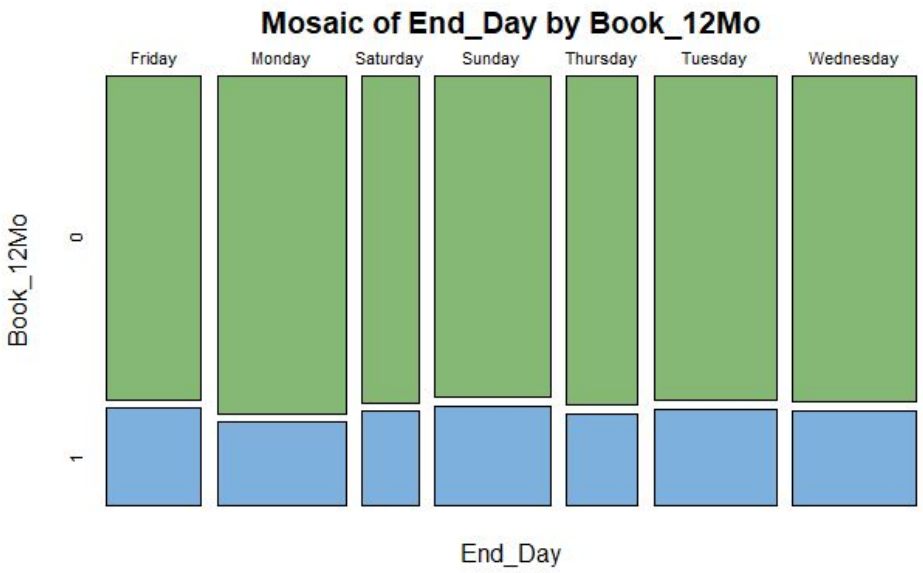
Metric	Statistic
N	23,459
Missing	0
Distinct Value	19
25th Percentile	11
50th Percentile	12
Mean	12.21
75th Percentile	14



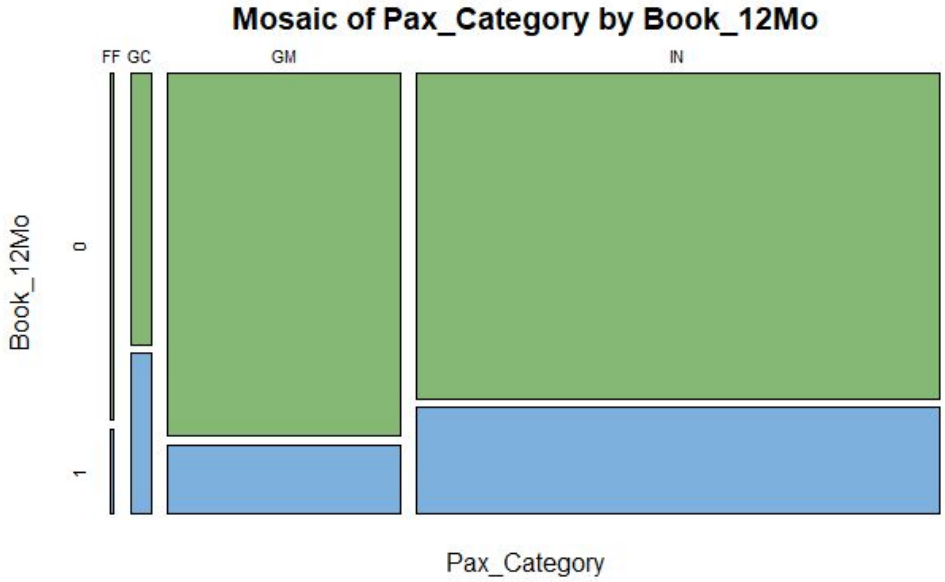
- **Start_Day:** Start_day is the day of the week of the starting day of the tour. Over 1 in 4 customers' starting day of their tour was on Thursday. It doesn't seem like there is a major difference in the likelihood of rebooking based on the start day.



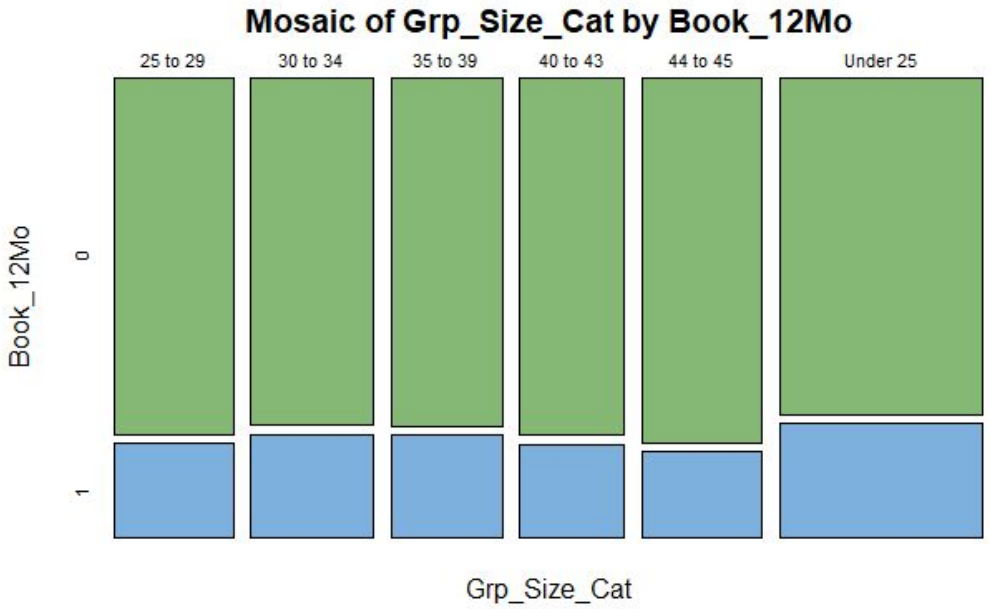
- **End_Day:** The End_Day is the day of the week of the last day of the tour. There are 23,459 values, 0 missing values, and 7 distinct values as expected. Customers were least likely to end their tour on a Saturday. It doesn't seem like there is a major difference in the likelihood of rebooking based on the day of the week they end the tour.



- **Pax_Category:** It appears that the majority of customers took individual traveler tours. There were no missing values for this variable. It seems based on the mosaic plot below that customers who were individual travelers (IN) and especially customers who were group coordinators (GC) most likely to book another tour in the next 12 months.

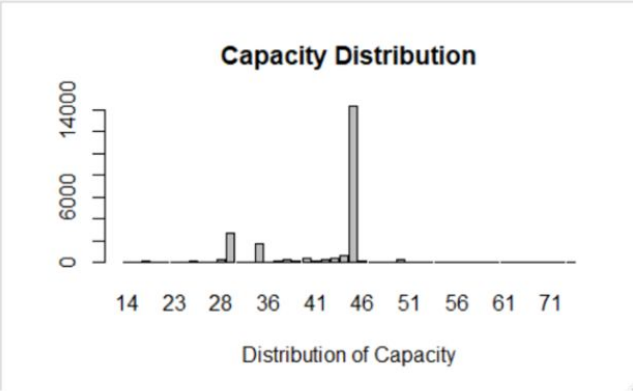


- **Grp_Size_Cat:** This variable represents the total number of people on the tour categorized by the company. There are 23,459 values, 0 missing values, and 6 distinct values. It does appear that the last category was mis-labeled (it overlaps with the previous category) so we renamed ’44 to 45’ instead of ’43 to 45’ based on the actual group size. It does not appear that this variable is strongly related to a customer booking another tour in the next 12 months based on the mosaic plot below.



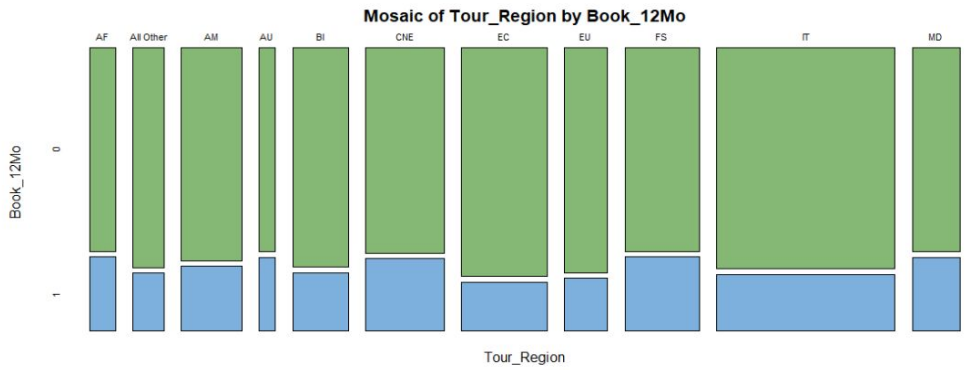
- **Grp_Size:** This variable represents the total number of people on the tour. It is continuous. The distribution appears slightly left-skewed, but the mean and the median are roughly equal (32 customers). The point-biserial correlation is 0.047, indicating a weak relationship between this variable and the target. This is the variable we used to inform the change the group size category variable above.
- **Capacity:** This variable represents the capacity of the tour. It is a continuous variable, but is probably more appropriate as a categorical variable based on the distribution shown below (see section on generated variables). Most of the values bunch up around 45, 30, and 35. Capacity does not appear to be strongly related to the target, judging by the mosaic plot below.

Metric	Statistic
N	23,459
Missing	0
Distinct Values	48
25th Percentile	39
50th Percentile	45
Mean	41.9
75th Percentile	45

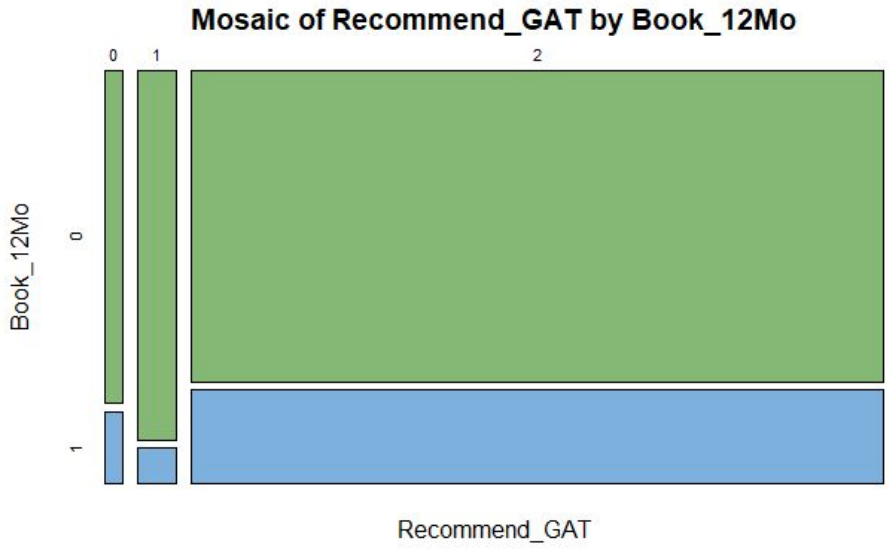


- **Tour_Region:** This variable represents the primary location that the tour went to. There are N = 23,459 values, missing values = 0, and 22 distinct values. Below are the expected values based on the data dictionary. The documented values make up approximately 77% of the data. We also kept EC and EU, which are not documented in the data dictionary, as their own categories and then grouped all the remaining distinct values into an “All Others” category, which represents a bit less than 5% of the final dataset.

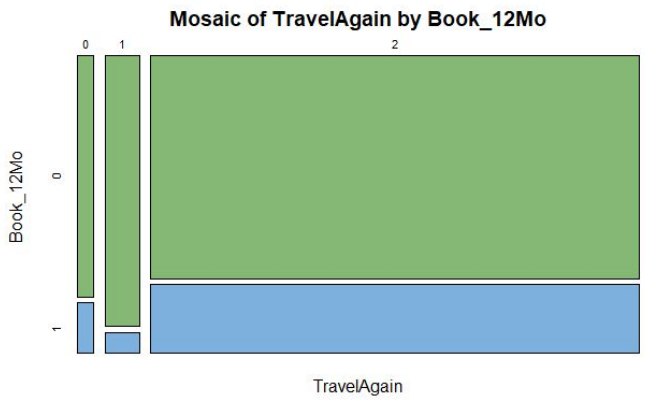
Category	Label	Count	Proportion
IT	Italy	5,989	25.5%
BI	Britain & Ireland	1,866	7.9%
CNE	Central & Northern Europe	2,644	11.3%
FS	France & Spain	2,500	10.7%
MD	Mediterranean Combos	1,612	6.8%
AM	Americas	2,020	8.6%
AF	Africa	887	3.8%
AU	Asian/AUN	551	2.4%
Total			77%



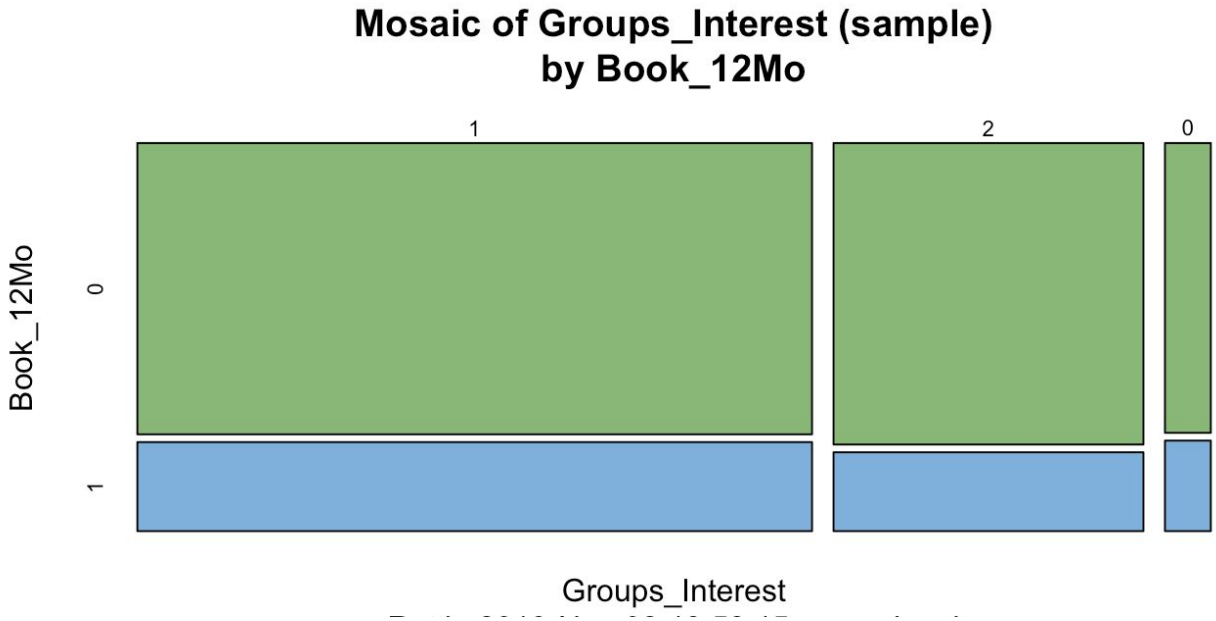
- **Recommend_GAT:** This variable is the response to the question – Would you recommend the company to your friends and family? 1 – No, 2 – Yes. There are 23,450 values, 0 missing values, and 3 distinct values. We only expected two distinct values from the documentation, but it appears some customers answered ‘0’. We did not change this 0 value. It does appear, as expected, that customers who would recommend the company to friends and family are more likely to book another tour within the next 12 months.



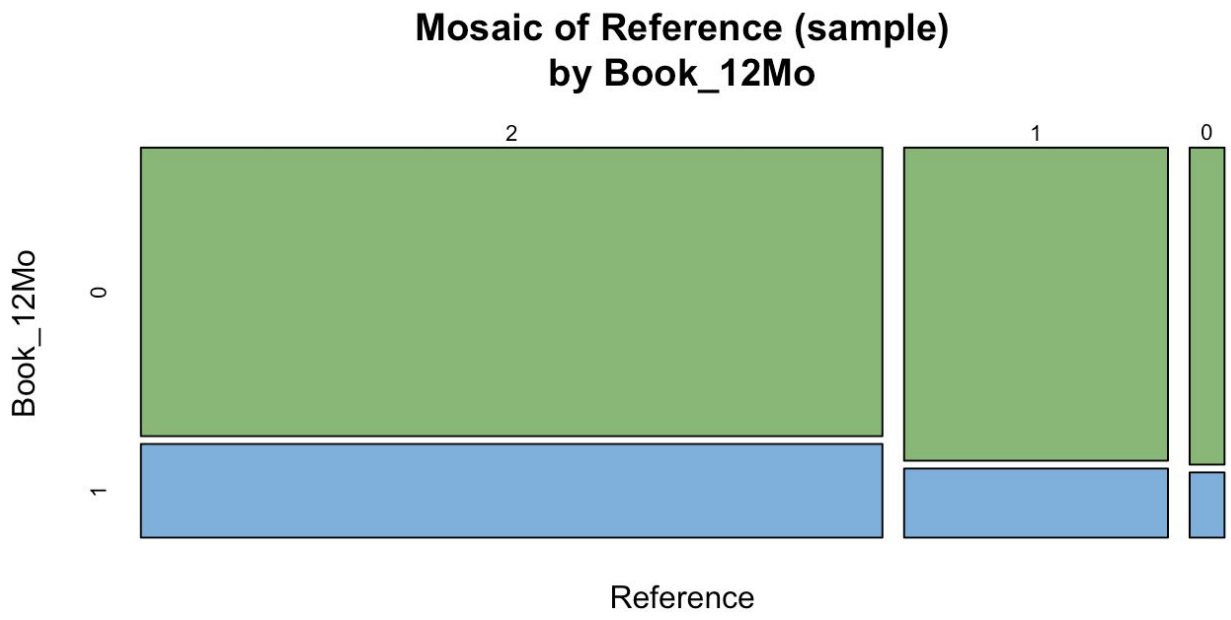
- **Travel_Again:** This variable is the response to the question - Would you take another tour with this company? 1 – No, 2 – Yes. Some customers answered ‘0’ which is an unexpected value for this variable. This variable and the target appear somewhat related based on the mosaic plot. Those who answered “yes” were significantly more likely to rebook within 12 months.



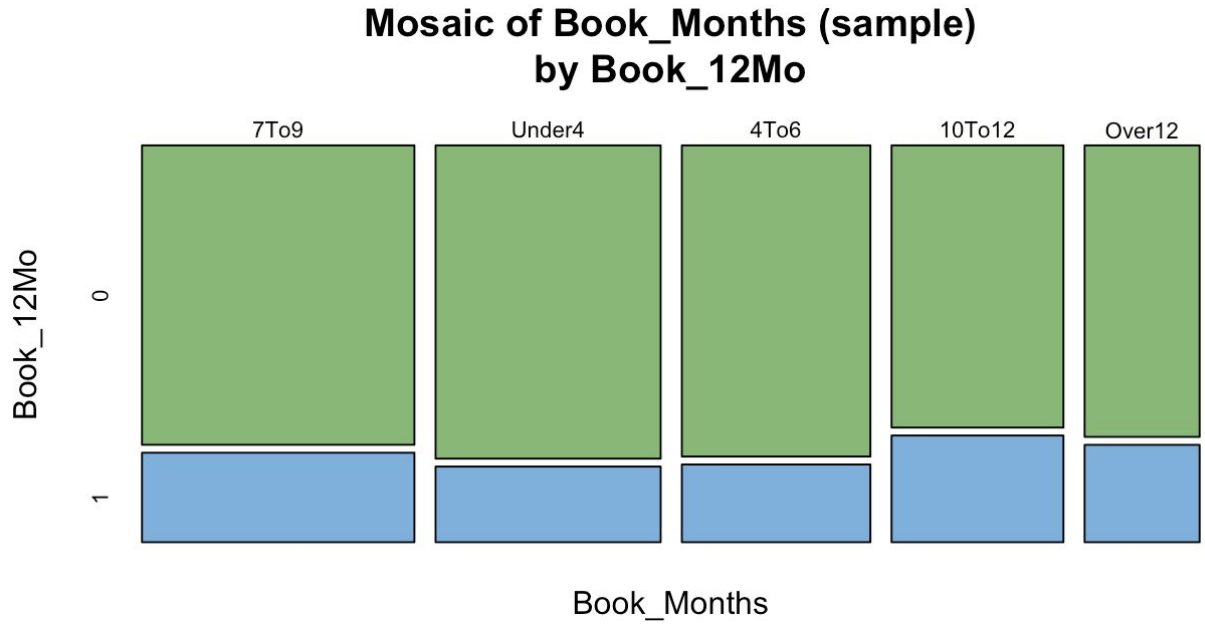
- **Groups_Interest:** This variable represents Response to the question – Would you be interested in hearing more about our group programs? 1 – No, 2 – Yes. Again, some customers answered ‘0’ but we left these 0 values. This variable does not appear to be related to the target variable based on the mosaic plot below.



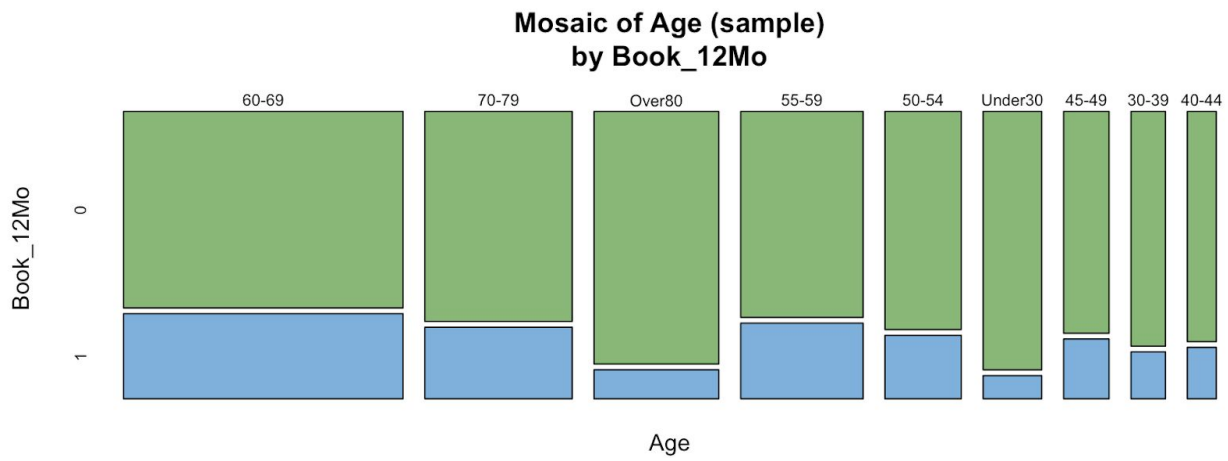
- **Reference:** This variable represents the response to question – May we use you as a reference for other travelers interested in taking this tour? 1 – No, 2 – Yes. There are some ‘0’ values but we did not change these to missing values. It appears that customers who respond that they cannot be used as a reference are less likely to book another tour in the next 12 months judging by the mosaic plot below.



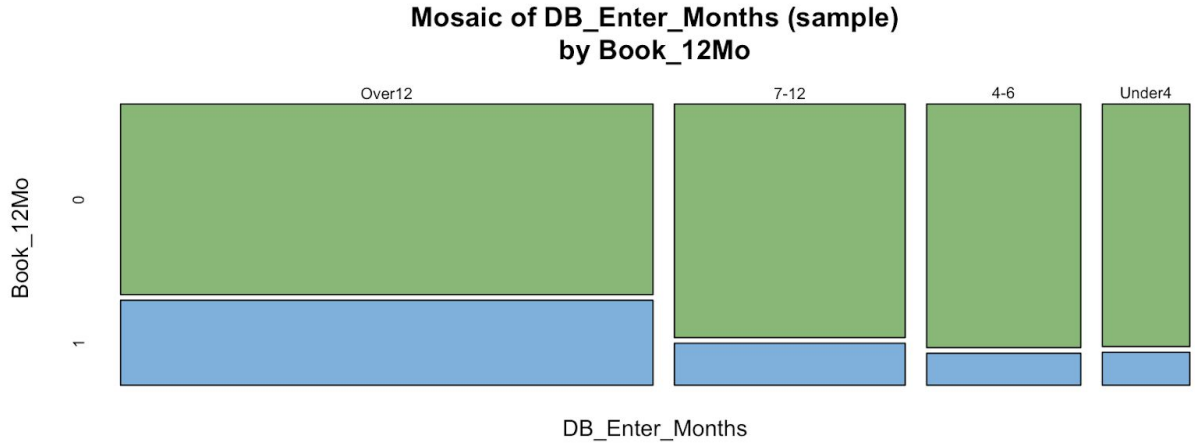
- **Book_Months:** The number of months prior to departure that the customer has made the reservation – For example, customer 123 booked the tour 3 months before he/she went on tour. This variable measures relatively highly on variable importance measures. As perhaps expected, it seems that those who book further further in advance are more likely to rebook within 12 months.



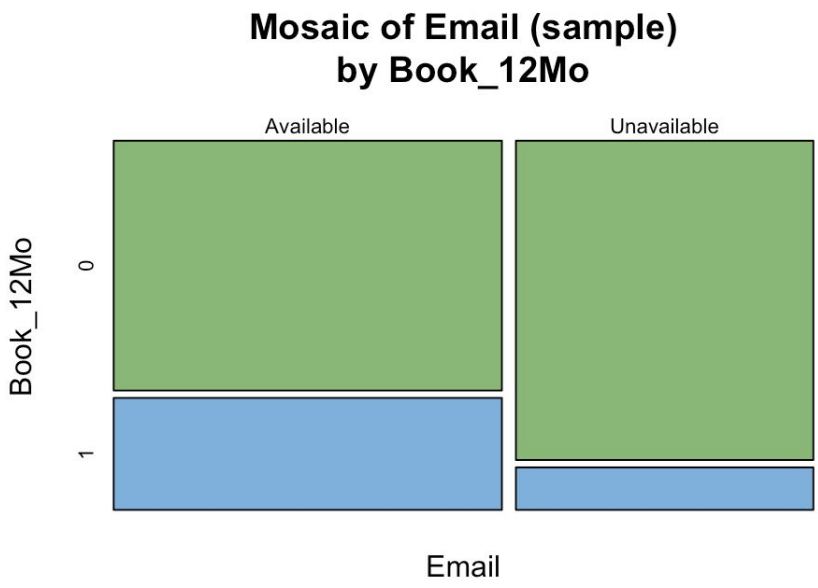
- **Age:** Age of the customer at the time of the tour. This variable ranks highly on the measures of variable importance.



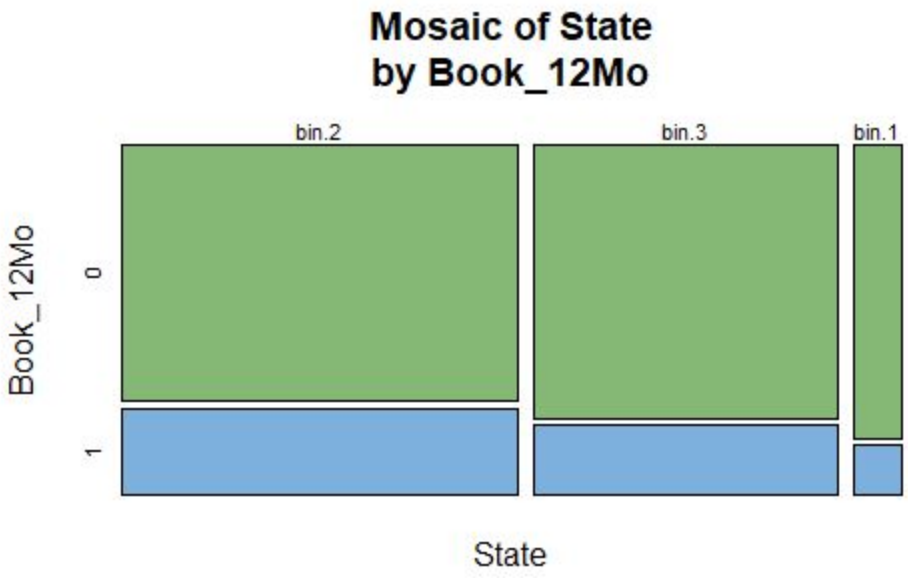
- **DB_Enter_Months:** The number of months prior to reservation that the customer has entered the database – For example, customer 123 booked the tour 3 months after he/she first requested the catalog. This variable ranks highly on measures of variable importance. It seems that those who enter the database earlier are more likely to rebook within 12 months.



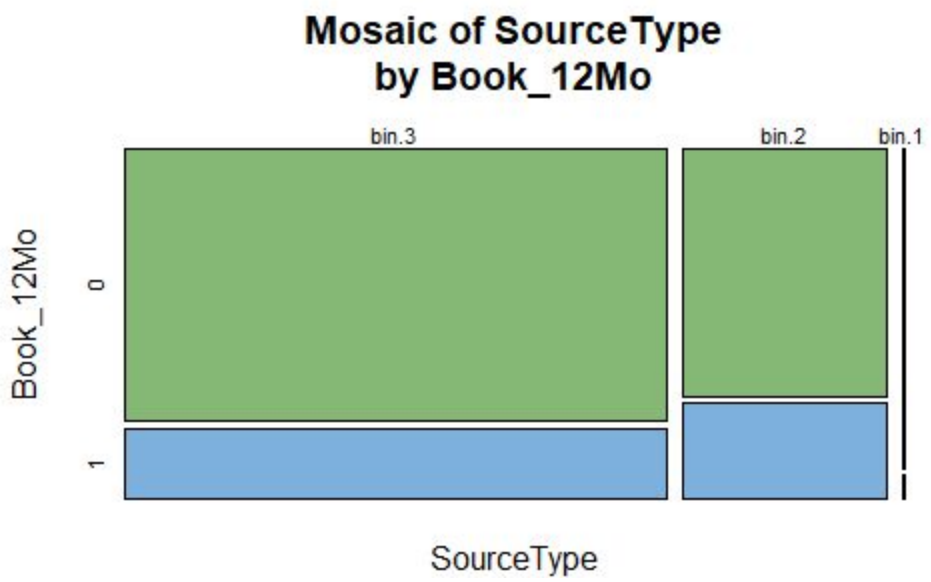
- **Email:** Indicates whether the email is available for this customer or not. Interestingly, this variable ranks near the top of the variable importance measures. This could be because those people receive email promotions which lead them to rebook or because those who are willing to give their email addresses in the first place are more likely to rebook or some combination of the two.



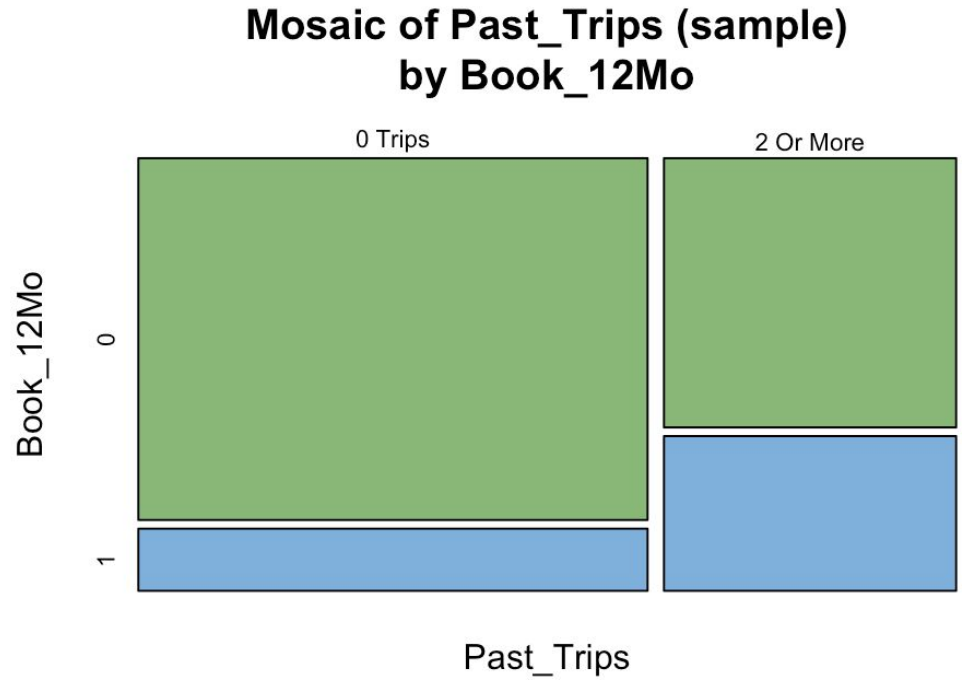
- **State:** The geographic location of the customer -- Derived variables could also be geographic areas (east, west, midwest, etc.) or location categories (Bible Belt etc.) This can be useful in finding out where the customers are located and how to better promote the company's services within the area/region.



- **Source Type**: The marketing category that brought in this customer into the database, can be internet, referral, organic web, or other. This allows the company to see which area of marketing works best and use it to optimize resources.

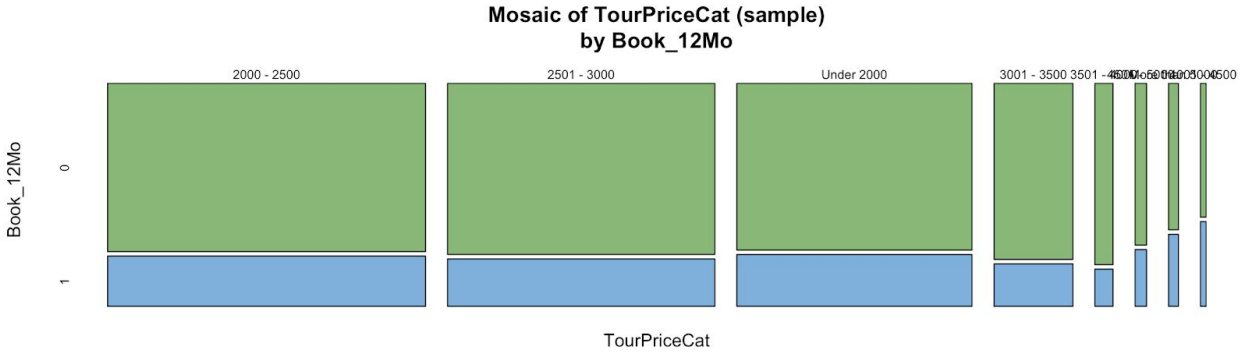


- **Past Trips**: The number of trips taken by the customer in the past. This can prove useful, seeing as if someone has purchased a trip (prior to their current trip), they might be enjoying the experience enough to book again. Indeed, according to the ROC Index, this is the most significant variable in determining rebooking, which can be seen in the mosaic plot below.

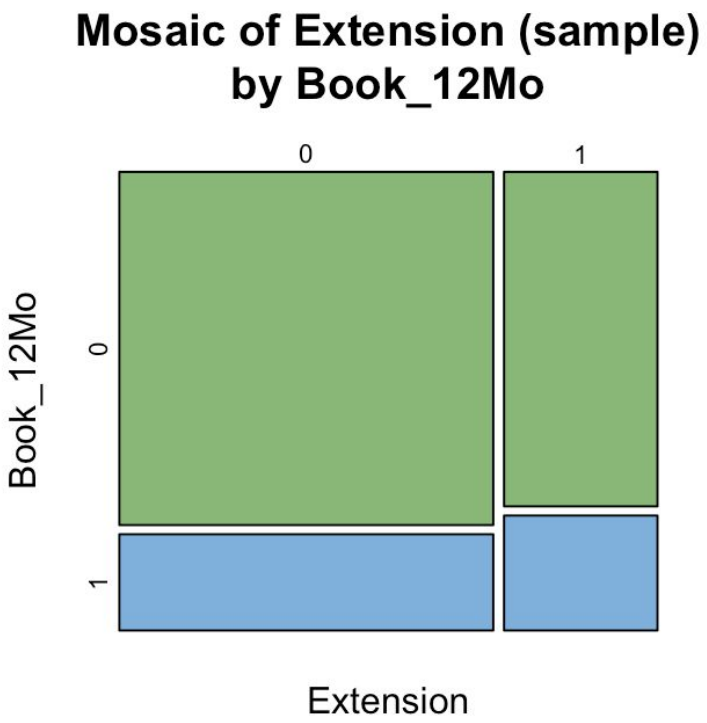


- **TourPriceCat**: The price of the tour that the customer bought. This allows the company to scale the occupancy for the tours based on what price range is frequently purchased. Using this as guidance for the amount of trips being

frequently taken can help bring in revenue and lessen costs. Although there are relatively few trips above $5000, customers from these trips tend to be more likely to rebook within 12 months, perhaps because they have more money to spend or because the more expensive trips are more enjoyable.

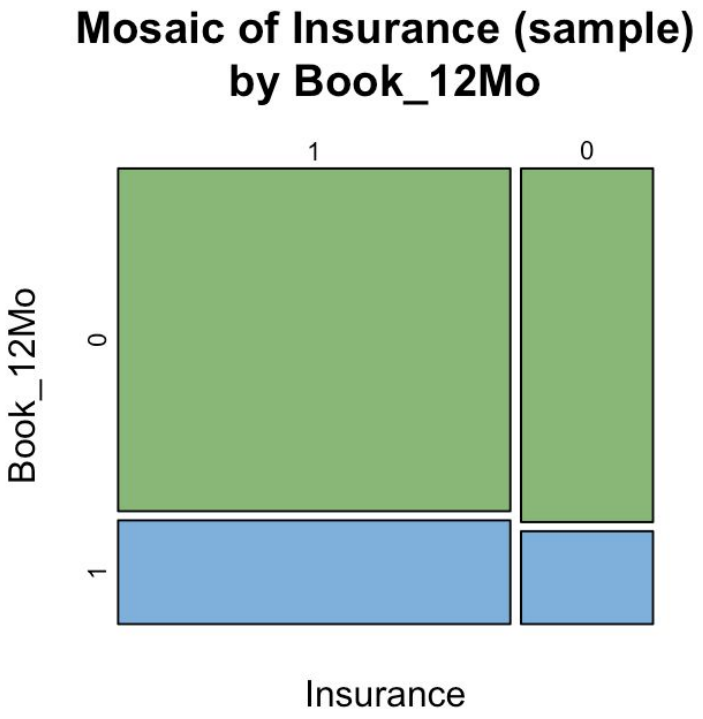


- **Extension**: A flag indicating whether the customer purchased the trip extension or not. We can use this information to target individuals that likely had a great experience and find insights as to what made the trip worth extending. Based on the mosaic plot, it doesn't seem like this variable is particularly correlated with rebookings within 12 months.



- **Insurance**: A flag indicating whether the customer purchased insurance or not. This can be paired with other variables to see the demographics of customers that are likely to purchase insurance and use it as a resource for marketing such

as discount codes or promotions. It doesn't seem like this variable is particularly strongly related to rebookings within a year, at least according to the mosaic plot.

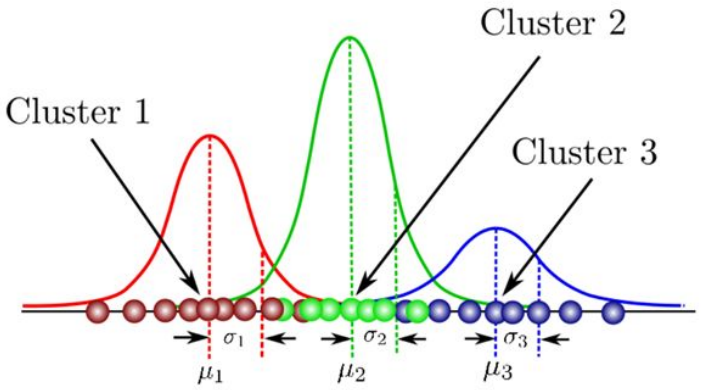


Section 3 - Gaussian Mixture Models for Clustering

Prior to our predictive modeling exercises that follow, we wanted to explore clustering the dataset to see if we could uncover relationships we wouldn't otherwise find and that might help improve the performance of our predictive models. Naturally, we began with k-means clustering. Recall that k-means clustering is a data mining technique for the task of identifying subgroups in a dataset that are similar. Data points that get assigned to another cluster are supposed to be different. The technique assumes that the clusters are spherical and that the clusters are of similar size.

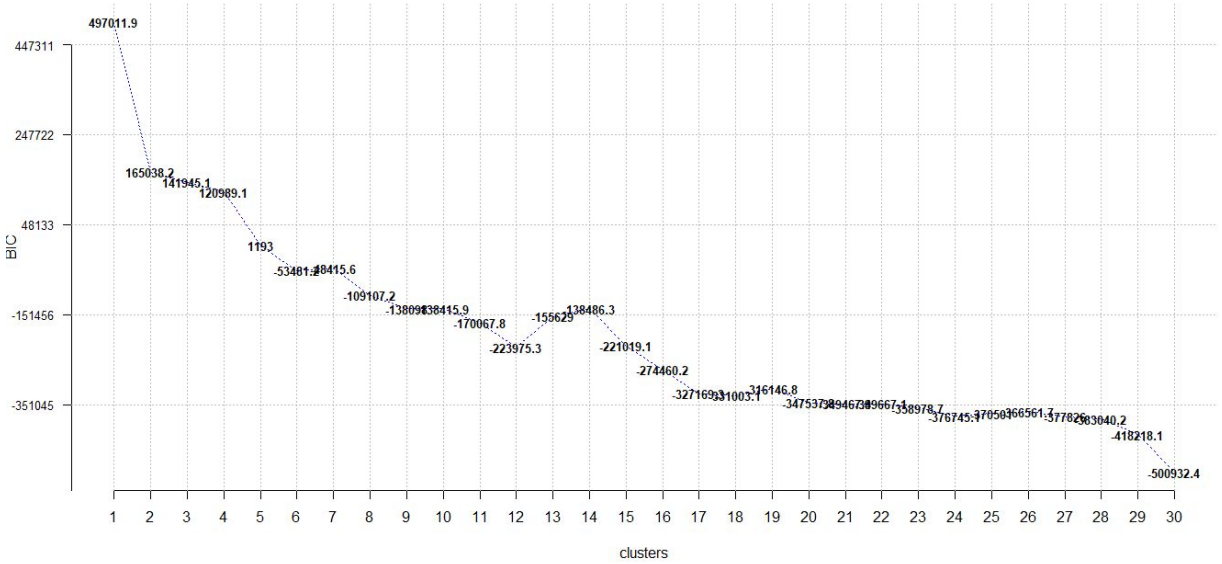
In addition, K-means clustering is a "hard" clustering method that assigns each data point to a cluster without any measure of certainty. In contrast, we explored the Gaussian Mixture Modeling clustering technique, a "soft" clustering method that assigns each data point a probability that that particular data point belongs to any of the clusters. We found that when trying to fit k-means clustering to some of the numerical data that the assumptions seem to have been violated (i.e. the clusters were not spherical and the clusters were not of similar size).

A Gaussian Mixture is a function that is comprised of several Gaussians. Each Gaussian has 3 parameters - the cluster mean, covariance to define width, and a mixing probability that defines the size of the Gaussian function. Below is a visual example:



In our first iteration of clustering prior to the class presentations, we fit GMM clustering to all of the variables in the dataset except the ID variables. We responded to the feedback we received by making adjustments. In our dataset the set of inputs relating to customer survey data were used to fit a Gaussian Mixture Model to create a classification cluster of three groups. In order to correct for the class imbalance issue in the data, we fit the GMM clustering algorithm using the downsampled inputted data.

In order to select the optimal number of clusters to use when fitting the Gaussian Mixture Model, we use the Elbow Method heuristic. We use BIC as our metric, although results were similar with other metrics. See below:



We select 3 clusters based on the heuristic. Below, you can see our tabulations of cluster sizes and the proportion of the customers in each cluster that booked another tour with the next 12 months in data that was downsampled.

```
     0    1
1 1015  349
2 4065 2176
3 6649 2727
```

```
           0         1
1 0.7441349 0.2558651
2 0.6513379 0.3486621
3 0.7091510 0.2908490
```

```
            0          1
1 0.05977269 0.02055238
2 0.23938520 0.12814322
3 0.39155527 0.16059125
```

And below you can find see our tabulations of cluster sizes and the proportion of the customers in each cluster that booked another tour with the next 12 months in data that was downsampled.

```
      0     1
1  1546   349
2  6383  2176
3 10278  2727
```

```
          0         1
1 0.8158311 0.1841689
2 0.7457647 0.2542353
3 0.7903114 0.2096886
```

```
           0          1
1 0.06590221 0.01487702
2 0.27209173 0.09275758
3 0.43812609 0.11624536
```

Moving forward in the modeling sections, we incorporate the cluster assignments into modeling since the clusters scored in the top quarter of our variable importance metrics. So it seemed worth our while to continue using them.

We responded to the great feedback we received during the presentation by narrowing the set of variables considered for the clustering exercises. Given more time, we would also improve our model by taking advantage of the probability distribution over the 11 clusters that the GMM algorithm provided for us. This was another helpful feedback we received from the class.

Section 4 - Model Building

Logistic Regression

The first step was to get rid of variables that we had created that caused errors due to lack of variance. These variables would be those that had maybe 2 categories, but only 1 or 2 observations in one category, while a couple thousand in the other. After that, I began to find out which model building function created a better model based on the AIC score. Doing this, I used R's backwards, forwards, and both iterations of logistic regression (the highest AIC came from the both iteration). Initially, I got an Fscore of about 30%, which seemed a bit low, so I decided to alter the threshold as an attempt to increase this. During this process, I also incorporated our downsampled data in order to test whether a smaller data set would help increase the Fscore or

lower the misclassification rate. Despite a slightly higher Fscore (about 1% increase), I decided that it wouldn't be worth the loss of information.

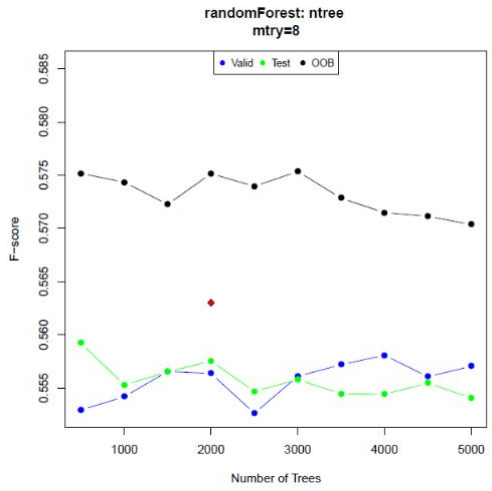
The final model for logistic regression came with an Fscore of nearly 43%, a misclassification rate of nearly 55%, about 75 variables, and an AIC of 10294. The most important variables are whether or not they have their email, are range, whether or not they had previously booked trips, and if they rated their optionals as excellent or not.

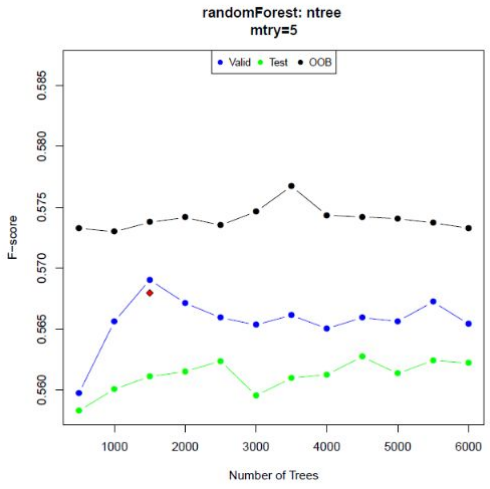
ANN

Using the final model from the logistic regression, we looked to create an ANN model. Here we tested the different amount of inputs, neurons, and hidden layers. We eventually settled on 100 inputs, 2 hidden layers with 100 and 20 neurons, and a hyperbolic tangent activation function. A threshold from ROC curve of 0.2089146 was used in classification resulting in an Fscore of about 50%.

Decision Tree/Random Forest

For our Random Forest (RF), hyperparameters parameters where selected to maximise a mean f-score with minimum variance by fixing a parameter and fitting models over a sequence of another. Iterations of parameter selection and optimisation tuned the model until the mean f-score seem to pato. For example, with mtry fixed at 8 the validation, training, and Out-of-Bag (oob) fscores for a sequence over ntree were plotted with the maximum mean fscore marked in red (shown below left).





The tuning process was concluded (shown above right) with m in five predictors considered at each split, for n in (500 to 2000) number of trees grown to maximum leaf size of one observation, and the best FR parameters were chosen at the maximum mean fscore. The final model was fit using the best parameters and importance measures. The Mean Decrease Accuracy Rank and Mean Decrease Gini Rank were calculated and their product was used to

sort the variables by decreasing importance. The final model parameters, scores, and five most impotent variables are shown below.

Final Random Forest Model

Hyperparameters:

- 1500 Number of Trees
- with 5 Predictors considered at each split.

Model Fscores: without importance and fited with importance

- Validation: 0.5632911, 0.5615764
- Test: 0.5586207, 0.5607735
- Out-of-Bag: 0.5720137, 0.5771996

Varable Importance: decreasing in the product of Mean Decrease Accuracy Rank and Mean Decrease Gini Rank

Input	MDA.Rank	MDG.Rank	Rank.Product
Past_Trips	68	67	4556
Email	66	66	4356
Tour_Region	67	65	4355
Age	63	68	4284
DB_Enter_Months	64	57	3648
Total_Return_Connect_Time	58	61	3538

Varable Selection

Below shows the variables that were removed before fitting and tunning the final RF model. These include the non-input and rejected variables such as Eval_ID, inputs used in Gaussian Mixture Model classification clusters, and four NA flag inputs. The four NA flag inputs were identified in prior RF modeling attempts and shown to be of little value. The Gaussian Mixture Model classification clusters will be described in the following section.

```
# > names(data.rf[-vars.rf])
# [1] "EvalID"                      "Cus_ID"                      "ProdTour_ID"
# [4] "SalesTourID"                 "Trip_no"                     "TourDate"
# [7] "Overall_Impression"          "Pre_Departure"               "Flight_Itin"
# [10] "TD_Overall"                  "Hotels_Avg"                  "Meals_Avg"
# [13] "GUSS_Avg"                    "Optionals_Avg"               "Bus_Avg"
# [16] "Hotel_2orUnder"              "Meals_2orUnder"              "GUSS_2orUnder"
# [19] "Optionals_2orUnder"          "Bus_2orUnder"                "Poor_Hotels"
# [22] "Poor_Meals"                  "Poor_GUSS"                   "Poor_Optionals"
# [25] "Poor_Buses"                  "Fair_Hotels"                 "Fair_Meals"
# [28] "Fair_GUSS"                   "Fair_Optionals"              "Fair_Buses"
# [31] "Good_Hotels"                 "Good_Meals"                  "Good_GUSS"
# [34] "Good_Optionals"              "Good_Buses"                  "Excellent_Hotels"
# [37] "Excellent_Meals"             "Excellent_GUSS"              "Excellent_Optionals"
# [40] "Excellent_Buses"             "HH_ID"                       "Promo_Disc"
# [43] "Optionals"                   "Domestic_Depart_Time"        "Intr_Arrival_Time"
# [46] "Intr_Depart_Time"            "Domestic_Arrival_Time"       "Tour_Region_na"
# [49] "Outbound_Domestic_Gateway_na" "Outbound_Intr_Gateway_na"    "Return_Domestic_Gateway_na"
# [52] "Return_Intr_Gateway_na"
```

Attempts to remove correlated input such as those transformed in phase I to define new inputs resulted in lower mean fscores. These inputs were kept in the final RF model.

```
##  [1] "Capacity"                    "Eval_Contact_Days"
##  [3] "Outbound_Connect_Time_Mins_1" "Outbound_Connect_Time_Mins_2"
##  [5] "Return_Connect_Time_Mins_1"   "Return_Connect_Time_Mins_2"
##  [7] "Outbound_Connect_Gateway1"    "Outbound_Connect_Gateway2"
##  [9] "Return_Connect_Gateway1"      "Return_Connect_Gateway2"
```

Section 5 - Model Revisions

As our final model ended up being our DT/RF model, we decided it would be best to focus on revising our complete model to see if we could modify something about our model or data in an attempt to increase our F-score, while keeping our variance low. Some changes were made to the phase I cleaning code to account for issues in scoring and fixing a key error. We accidentally miss labeling a named key for Past_Trips, recoding the factor of three levels into a binary variable of zero or two-or-more trips. Next we decided to drop some of the survey data variables by modeling them as a single classification cluster (see Section 3). Collectively, these revisions allowed our final model to be more concise, and provided a slightly higher mean fscore shown below.

	Without Importance	With Importance
Validation	0.5632911	0.5615764
Test	0.5586207	0.5607735
Out-of-Bag	0.5720137	0.5771996

Boosted from:

	Without Importance	With Importance
Validation	0.5385413	0.5427536
Test	0.5401309	0.534825
Out-of-Bag	0.5525292	0.5546703

Section 6 - Final Model

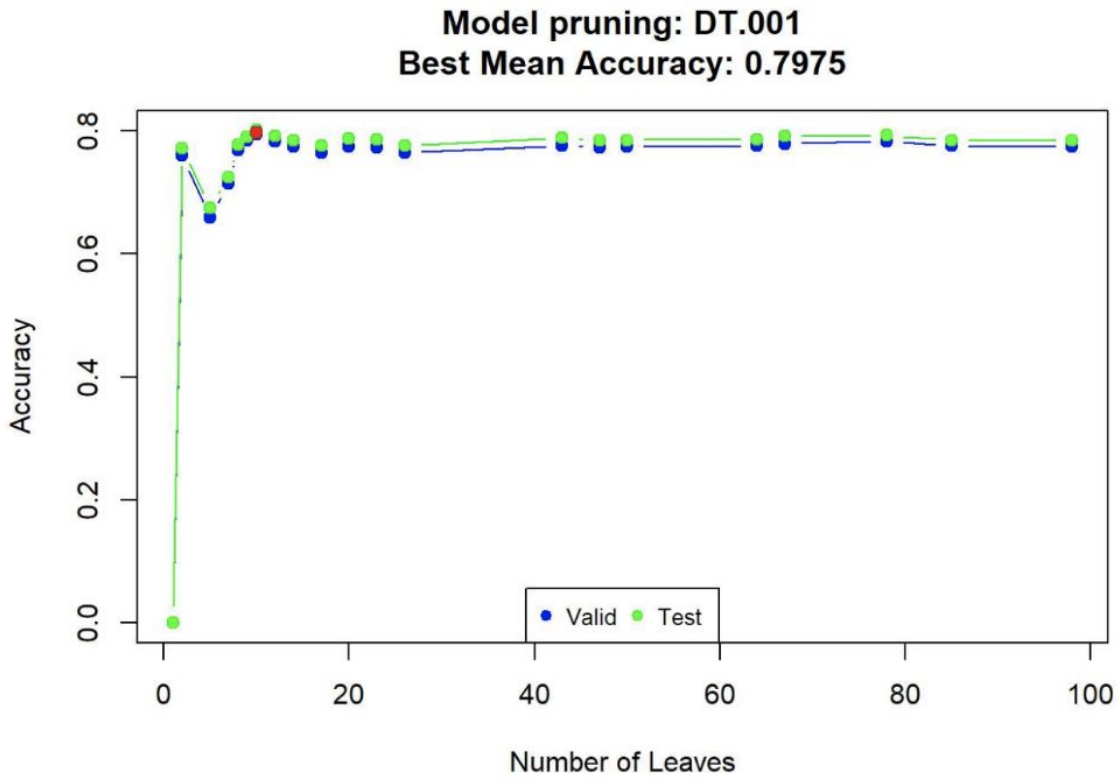
For our final model, we selected the one with the highest F-score out of the decision tree, logistic regression, artificial neural network, and random forest. We found that the random forest model performed the best: it consistently returned the highest F-score.

Section 6a - Full Model

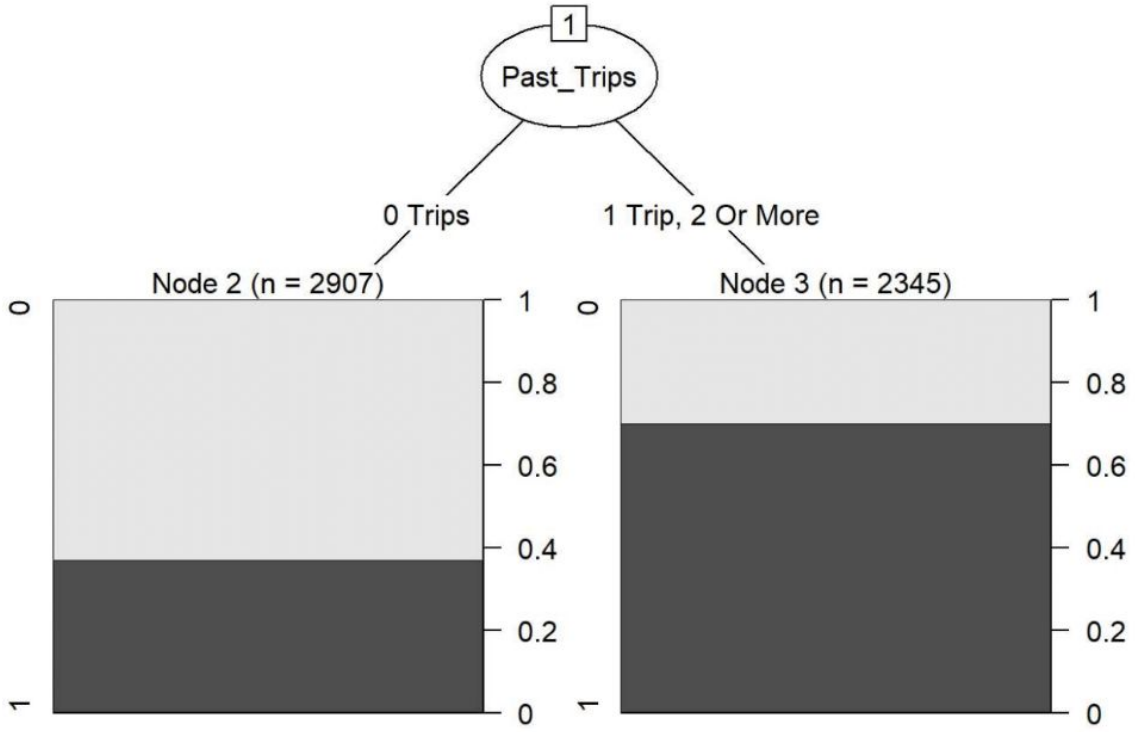
The full model is a random forest with 1500 trees and 5 predictors considered at each split. There are 86 total predictors used in the model. The process by which we arrived at these parameters was explained above in the model building section so I will not repeat that explanation here. The fact that the random forest model ended up performing the best was not particularly surprising, as this was consistent with the results we have seen when comparing different types of models in other contexts. The power of the random forest model comes from the fact that it essentially “polls” the crowd regarding the proper prediction. While decision trees offer simplicity and ease of computation, they can be very susceptible to small changes in the data. Random forests, while more computationally expensive and more difficult to interpret, handle this issue elegantly by growing many trees and therefore decreasing the importance of any given one.

Section 6b - Decision Tree Interpretation

One method of interpreting complex models is to use simpler models to capture the broad structure of the more complex model. In this case, we fit a decision tree to our final model. The target variable for the decision tree is not whether an individual re-books within 12 months. Instead, the target variable is the prediction made by the more complex model, in this case a random forest model. In other words, the decision tree is predicting the prediction, not the underlying target variable. By training the decision tree on the random forest model, we can use the decision tree to explain, in a general sense, how the random forest is making predictions. The plot below shows the pruning process for explanatory decision trees.



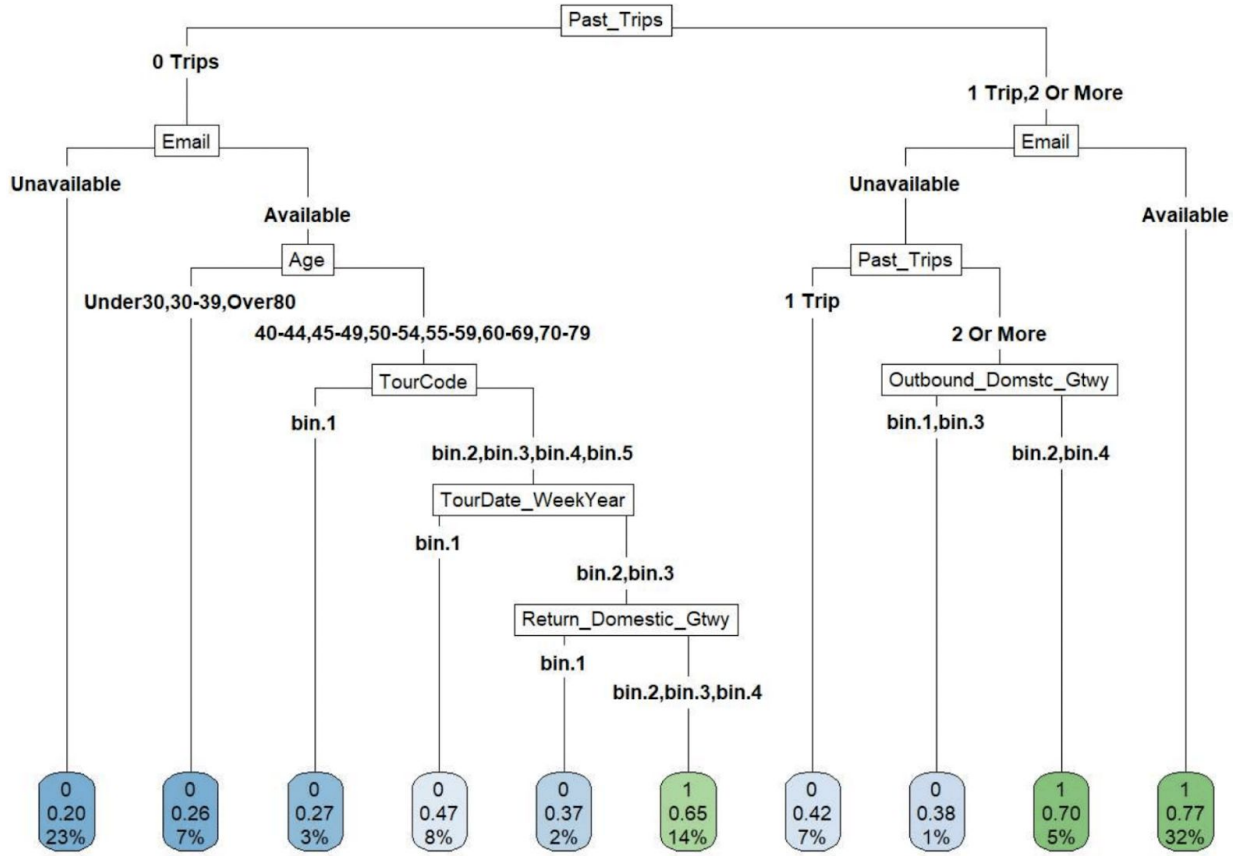
As the plot shows, while the highest accuracy occurs with 10 leaves, the accuracy doesn't increase markedly after the tree with two leaves. This tree with two leaves is based entirely on the variable Past_Trips, shown below.



It makes sense that this variable would be very helpful in predicting whether a customer is going to re-book within 12 months. If an individual has re-booked in the past, chances are that person is more likely to re-book again in the future than a person is traveling with the company for the first time. This insight could mean different things for the company depending on its perspective. On the one hand, it suggests the company should focus efforts on loyal customers because they are more likely to re-book, although in some sense this logic is circular: those who re-book are more likely to re-book. On the other hand, this insight may mean that the company should focus more resources on getting first-time travelers to re-book once and they will likely re-book again in the future. This insight--that those who've traveled on one or more trips in the past are more likely to re-book within 12 months--also doesn't help understand why these individuals chose to re-book or how to identify these frequent travelers.

The decision tree below includes 10 leaves. The colored shape at the bottom shows the prediction, the proportion of that leaf that have Book_12Mo equal to 1, and finally the proportion of the total that this leaf represents. As is shown by the decision tree, Past_Trips, Email, and Age have a significant effect on the final proportion of customers who book again. For instance, customers on their first trip who do not provide an email rebook only 20% of the time and compost 23% of the sample whereas customers who have gone on one or more trips and provide their email rebook 77% of the time and consist of 32% of the sample. This decision tree has approximately 80% accuracy, meaning it predicts the proper prediction of our random forest

about 80% of the time.



Section 6c - Variables of Interest

There are several variables worth exploring further, either because they have a particularly large effect on whether customers re-book within 12 months, or because they are interesting in their own right. The list below shows the dozen variables with the most predictive power in the random forest model for whether customers re-booked within 12 months. The relevance of Past_Trips is explained above but it's worth discussing several others.

Per the list below, the variable Email, while seemingly trivial, can be considered a proxy for interest in future trips. Those customers who provide an email address are more likely to be loyal customers than those who do not. The decision of whether to provide an email address is probably based on many factors, similar factors to those the customers consider when deciding whether to re-book. In addition, assuming that the tour company sends out emails with offers and promotions, it could be that these emails themselves make an individual more likely to re-book. Selection bias makes it difficult to disentangle which of these effects, or both, are occuring. It may be prudent for the company to perform some A/B testing to see if it is the emails themselves that are making customers more likely to re-book or if it's the simple fact that the customer was willing to provide his or her email address.

Age may be a particularly helpful variable for the company to consider because it can target individuals of a certain age if it knows that these individuals are more likely to re-book than individuals in another age group. Interestingly, many of the variables most helpful for predicting whether someone is going to re-book have to do with attributes of the tour itself, instead of a customer's subjective assessment of the tour's components (food, hotels, etc). There are lots of reasons for this, some of which involve the granularity of the rating variables (i.e. there are many of them, each measuring a very specific assessment), but one insight may also be that the company shouldn't read too deeply into customer's subjective ratings of their trip experience, at least in so far as it affects their likelihood to re-book.

The connect time variables are rated fairly highly. One reason for this may relate to the psychological principle that people are generally more likely to remember the beginning and the end of an experience instead of the middle. Assuming the customer does not live near the location of the tour, each tour begins and ends with air travel. This insight may be reason for making sure that each customer's travel experience goes smoothly.

Interestingly, the only variables out of the ten that relate to the tour itself are the tour region, the end day (the day of the week on which the tour ended), and the tour code. The clusters that capture subjective ratings of the customers' experience on the trip is not included in the top ten--in fact, it's not even in the top 20, suggesting the company should likely avoid targeting customers who said they had very positive trip experiences.

Input	MDA.Rank	MDG.Rank	Rank.Product
Past_Trips	68	67	4556
Email	66	66	4356
Tour_Region	67	65	4355
Age	63	68	4284
DB_Enter_Months	64	57	3648
Total_Return_Connect_Time	58	61	3538
TourCode	65	54	3510
End_Day	59	59	3481
Total_Outbound_Connect_Time	56	60	3360
Outbound_Connect_Time_Mins_1	52	62	3224

Section 7 - Conclusion

In conclusion, after exploring the dataset, cleaning the dataset, transforming and imputing variables, fitting and refining a variety of models and selecting a final model, we found that the random forest model performed the best with 86 variables, 1500 trees, and 5 variables at each split. We hope this model and the explanations we provide will help XYZ Tours target those customers most likely to rebook.

References

1. Gaussian Mixture Models Explained.

Appendices for Phase 1

Appendix A: Additional Variable Importance Measures for Categorical Variables

Variable Name	χ^2	χ^2 **Rank**	χ^2 **Simulated**	χ^2 **Simulated Rank**	**ROC**	**ROC Rank (Categorical Only)**
Past_Trips	2.55E-168	1	0.0005	13.5	0.647769	1
Email	3.00E-129	2	0.0005	13.5	0.632545	2
DB_Enter_Months	3.01E-108	3	0.0005	13.5	0.626542	3
Age	3.95E-79	4	0.0005	13.5	0.506196	29
Return_Domestic_Gateway	1.97E-54	5	0.0005	13.5	0.522624	13
Outbound_Domestic_Gatewa	2.66E-52	6	0.0005	13.5	0.522956	12

y						
Outbound_Intr_Gateway	1.20E-47	7	0.0005	13.5	0.580477	4
Return_Intr_Gateway	1.61E-45	8	0.0005	13.5	0.522072	15
Pax_Category	7.34E-33	9	0.0005	13.5	0.538262	7
State	1.88E-24	10	0.0005	13.5	0.518927	20
TravelAgain	2.23E-23	11	0.0005	13.5	0.531312	10
SourceType	1.18E-22	12	0.0005	13.5	0.544732	5
TourCode	7.96E-18	13	0.0005	13.5	0.536383	8
TourDate_WeekYear	7.09E-17	14	0.0005	13.5	0.502882	35
Recommend_GAT	2.44E-14	15	0.0005	13.5	0.521452	17
Reference	1.14E-13	16	0.0005	13.5	0.538506	6
Tour_Season	2.08E-11	17	0.0005	13.5	0.503908	33
Tour_Region	3.82E-10	18	0.0005	13.5	0.501242	37
Grp_Size_Cat	2.55E-08	19	0.0005	13.5	0.5216	16
Book_Months	3.7E-07	20	0.0005	13.5	0.532887	9
Extension	1.81E-06	21	0.0005	13.5	0.524135	11
Main_Ext	8.98E-06	22	0.0005	13.5	0.522311	14
Tour_Type	1.85E-05	23	0.0005	13.5	0.515317	24
Insurance	0.000104	24	0.0005	13.5	0.518641	21
Voucher_Event	0.00011	25	0.0005	13.5	0.505167	31
Groups_Interest	0.000361	26	0.001	27.5	0.521144	18
Domestic_Depart_Time_AM	0.000474	27	0.0005	13.5	0.519121	19
FY	0.00086	28	0.001	27.5	0.508106	26

Intr_Arrival_Time_AM	0.003666	29	0.004998	30	0.517937	22
FltGty	0.004626	30	0.002999	29	0.505808	30
Intr_Depart_Time_AM	0.027671	31	0.027986	31.5	0.516236	23
Start_Day	0.02897	32	0.027986	31.5	0.507607	27
TourPriceCat	0.04034	33	0.048476	34	0.507148	28
Capacity_Cat	0.050519	34	0.046477	33	0.513759	25
Contact_Event	0.05793	35	0.055972	35	0.503626	34
Complaint_Event	0.126975	36	0.114943	36	0.501443	36
End_Day	0.151511	37	0.153423	37	0.504229	32
Domestic_Arrival_Time_AM	0.471999	38	0.450775	38	0.500077	38
Promo_Disc	1	39	1	39	0.500055	39

Appendix B: Additional Variable Importance Measures for Numerical Variables

Variable Name	**t- test value**	**T test value Rank**	**ROC**	**ROC Rank (Numerical Only)**
Overall_Impression	1.65E-23	1	0.552665789	1
Pre_Departure	7.38E-20	2	0.54877636	2
Grp_Size	5.67E-09	3	0.536157307	3
Hotels_Avg	9.13E-06	4	0.527509041	6
Meals_Avg	1.27E-05	5	0.52312731	8
Excellent_Optionals	6.80E-05	6	0.528954532	5
Good_Buses	0.000148211	7	0.511060436	23
Excellent_GUSS	0.000208791	8	0.523271095	7

Excellent_Hotels	0.000297921	9	0.529584135	4
Capacity	0.000430252	10	0.52068496	9
Poor_Meals	0.000769314	11	0.51251559	21
GUSS_Avg	0.000793035	12	0.51332355	20
Good_Hotels	0.002300035	13	0.516789863	15
Flight_Itin	0.002547094	14	0.516944566	14
Tour_Days	0.003501268	15	0.518259886	12
Total_Outbound_Connect_Time	0.003600279	16	0.507160173	29
Meals_2orUnder	0.003889735	17	0.515968998	16
Excellent_Meals	0.006290942	18	0.515281336	18
Optionals_Avg	0.008460289	19	0.501363591	44
Return_Connections	0.016693056	20	0.51452495	19
Optionals	0.018349143	21	0.519978873	10
Outbound_Connect_Time_Mins_1	0.023808542	22	0.511918215	22
Hotel_2orUnder	0.032767767	23	0.515627154	17
GUSS_2orUnder	0.035305357	24	0.506744857	31
Outbound_Connect_Time_Mins_2	0.060551961	25	0.50340884	40
Poor_Hotels	0.070907639	26	0.508015191	28
Fair_GUSS	0.07204044	27	0.504798281	34
Good_Meals	0.096837668	28	0.506988341	30
Fair_Meals	0.108828351	29	0.510991942	24
Good_Optionals	0.128450434	30	0.504537647	35
Fair_Hotels	0.129711492	31	0.510510495	25
Poor_GUSS	0.157036281	32	0.503068314	41
Bus_2orUnder	0.211720358	33	0.501852964	42
Return_Connect_Time_Mins_2	0.24323173	34	0.519304994	11
Fair_Buses	0.276111234	35	0.501400818	43
Outbound_Connections	0.323260826	36	0.50599349	32
Excellent_Buses	0.32692905	37	0.508297304	27

Eval_Contact_Days	0.340605922	38	0.503582596	39
Fair_Optionals	0.503652002	39	0.504476556	36
Poor_Buses	0.538651559	40	0.500399024	47
Bus_Avg	0.558637802	41	0.517079442	13
Return_Connect_Time_Mins_1	0.561392074	42	0.510177895	26
Optionals_2orUnder	0.665192908	43	0.504028844	37
Good_GUSS	0.672763416	44	0.503776848	38
TD_Overall	0.709051537	45	0.500894713	46
Poor_Optionals	0.754289362	46	0.501001501	45
Total_Return_Connect_Time	0.771451226	47	0.505618037	33

Appendix C: Bins for Outbound_Domestic_Gateway

	Outbound_Domestic_Gateway	**Categories**		**Outbound_Domestic_Gateway**	**Categories**
1	ABE	bin.1	106	NYC	bin.2
2	ABI	bin.1	107	ORD	bin.2
3	ABY	bin.1	108	PFN	bin.2
4	AMA	bin.1	109	PHL	bin.2
5	ATW	bin.1	110	PIT	bin.2

6	BDL	bin.1	111	RDU	bin.2
7	BGM	bin.1	112	RIC	bin.2
8	BGR	bin.1	113	ROC	bin.2
9	BIS	bin.1	114	SAN	bin.2
10	BMI	bin.1	115	SAT	bin.2
11	BOI	bin.1	116	SDF	bin.2
12	BUR	bin.1	117	SEA	bin.2
13	BZN	bin.1	118	SFO	bin.2
14	CAK	bin.1	119	SJT	bin.2
15	CHA	bin.1	120	SMF	bin.2
16	CHO	bin.1	121	SRQ	bin.2

17	CLE	bin.1	122	SYR	bin.2
18	COS	bin.1	123	TLH	bin.2
19	DAB	bin.1	124	TPA	bin.2
20	DAY	bin.1	125	TUS	bin.2
21	DRO	bin.1	126	TYS	bin.2
22	EWN	bin.1	127	VPS	bin.2
23	FLL	bin.1	128	WAS	bin.2
24	FSD	bin.1	129	XNA	bin.2
25	FWA	bin.1	130	YOW	bin.2
26	GEG	bin.1	131	YYZ	bin.2
27	GFK	bin.1	132	ANC	bin.3

28	GJT	bin.1	133	ATL	bin.3
29	GNV	bin.1	134	BNA	bin.3
30	GTF	bin.1	135	BTV	bin.3
31	HNL	bin.1	136	CID	bin.3
32	ILM	bin.1	137	CRW	bin.3
33	JAN	bin.1	138	DAL	bin.3
34	L01	bin.1	139	DEN	bin.3
35	LEX	bin.1	140	DLH	bin.3
36	LFT	bin.1	141	DTW	bin.3
37	LON	bin.1	142	EVV	bin.3
38	LYH	bin.1	143	FAR	bin.3

39	MCI	bin.1	144	FMY	bin.3
40	MEM	bin.1	145	FNT	bin.3
41	MFR	bin.1	146	GSO	bin.3
42	MGM	bin.1	147	GSP	bin.3
43	MLI	bin.1	148	HOU	bin.3
44	MLU	bin.1	149	HSV	bin.3
45	MOT	bin.1	150	LBB	bin.3
46	MQT	bin.1	151	LIT	bin.3
47	MSY	bin.1	152	LO1	bin.3
48	OAK	bin.1	153	MOB	bin.3
49	ORL	bin.1	154	MSN	bin.3

50	PAR	bin.1	155	MYR	bin.3
51	PDX	bin.1	156	OKC	bin.3
52	PIA	bin.1	157	OMA	bin.3
53	PIR	bin.1	158	ONT	bin.3
54	PLN	bin.1	159	ORF	bin.3
55	RAP	bin.1	160	PBI	bin.3
56	RNO	bin.1	161	PHX	bin.3
57	ROA	bin.1	162	PWM	bin.3
58	SAV	bin.1	163	SJU	bin.3
59	SBN	bin.1	164	SLC	bin.3
60	SCE	bin.1	165	STL	bin.3

61	SGF	bin.1	166	AGS	bin.4
62	SJC	bin.1	167	ALO	bin.4
63	SNA	bin.1	168	AUS	bin.4
64	SPI	bin.1	169	AVL	bin.4
65	STC	bin.1	170	AZO	bin.4
66	STT	bin.1	171	BIL	bin.4
67	TOL	bin.1	172	BQK	bin.4
68	TRI	bin.1	173	CHS	bin.4
69	TUL	bin.1	174	CLD	bin.4
70	TUP	bin.1	175	CMI	bin.4
71	YHZ	bin.1	176	CRP	bin.4

72	YLW	bin.1	177	CSG	bin.4
73	YQR	bin.1	178	EUG	bin.4
74	YVR	bin.1	179	FAT	bin.4
75	YWG	bin.1	180	HAR	bin.4
76	YXE	bin.1	181	HRL	bin.4
77	YYC	bin.1	182	LAN	bin.4
78	ABQ	bin.2	183	MAF	bin.4
79	ALB	bin.2	184	MBS	bin.4
80	BHM	bin.2	185	MEI	bin.4
81	BOS	bin.2	186	MFE	bin.4
82	BTR	bin.2	187	MKG	bin.4

83	BUF	bin.2	188	MLB	bin.4
84	CAE	bin.2	189	MSO	bin.4
85	CLT	bin.2	190	PNS	bin.4
86	CMH	bin.2	191	PSP	bin.4
87	CPR	bin.2	192	PVD	bin.4
88	CVG	bin.2	193	RHI	bin.4
89	DFW	bin.2	194	SBA	bin.4
90	DSM	bin.2	195	SBP	bin.4
91	ELP	bin.2	196	SHV	bin.4
92	GPT	bin.2	197	TVC	bin.4
93	GRB	bin.2	198	YEG	bin.4

94	GRR	bin.2	199	YFC	bin.4
95	ICT	bin.2	200	YMQ	bin.4
96	IND	bin.2	201	YUL	bin.4
97	JAX	bin.2	202	ERI	bin.1
98	LAS	bin.2	203	CWA	bin.1
99	LAX	bin.2	204	MHT	bin.1
100	LNK	bin.2	205	C	bin.1
101	MCO	bin.2	206	RDM	bin.1
102	MIA	bin.2	207	AEX	bin.1
103	MKE	bin.2	208	FAY	bin.1
104	MSP	bin.2	209	MRY	bin.1

105	ABE	bin.1	210	SWF	bin.1

Comment Summary

Page 35

1. That is quite a few inputs used in a logistic regression model.

2. How exactly is the mean f-score calculated? Is it the average of f-scores over OOB, validation and testing?
3. What strategy was used in the random forest to fight class imbalance?